LINKS IN SCIENCE

LINKS IN SCIENCE

Linking Network and Bibliometric Analyses in the Study of Research Performance

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof. mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 11 december 2012
klokke 16:15 uur

door

Clara Maria Calero Medina

geboren te Madrid, Spanje
in 1973

Promotiecommissie

Promotor:	Prof. dr. A.F.J. van Raan
Co-promotor:	Dr. E.C.M. Noijons
Overige leden:	Prof. dr. S. Hornbostel (Humboldt University Berlin) Prof. dr. J. Reedijk Prof. dr. mr. C.J.J.M. Stolker Prof. dr. P.F. Wouters

To my uncle

Cover: *Redes*
Photography and Design by *Alicia Fernández Solla*.

Caminante, no hay camino,
se hace camino al andar.

(Traveler, there is no road;
you make your own path as you walk)

Antonio Machado (1875-1939)
(Translated by Mary Berg & Dennis Maloney)

Acknowledgements

This was a long walk but luckily I did not walk by myself the whole time. There has been a great deal of people that accompanied me in one way or another. If I look back you helped me to choose paths. You made the difference. I cannot mention all of you here but I want to thank you all. This walk was much nicer because all of you.

Thanks Ton for giving me that opportunity of working at CWTS in 2003. I started to work with Ed and Martijn. Ed was the first member of CWTS whom I met and with whom I came to work initially at CWTS. Ed your way of listening and thinking is something that I will always appreciate. Thanks Ed. Martijn patiently taught me to work with the software that he had created. He has become a great support and friend during these years. Thank you Martijn. Later on I started to work with Henk. He was always there with his wise words, teachings and support. Thanks Henk. In the mean time I started to work more and more with Thed. Working in projects together with him I have learnt a lot. Thed, thank you for always being there. Rodrigo arrived later on. His energy and enthusiasm for bibliometrics is in a way contagious. Thanks Rodrigo for being a good colleague and a good friend.

During these years at CWTS I would not have survived without the help of Suze, Christine, and later on Anne Marie. Thanks very much to the three of you. I also want to thanks Maria. She was there making sure I was taking care of myself when I needed to rest my energy. Bert, thanks for being a great office mate during several years. Peter and Erik, thanks for helping me with my stupid questions every time I asked you for. Many thanks also to the rest of colleagues at CWTS. Many of you came not so long ago and you brought a breath of fresh air. It is nice.

Special thanks go to my Professor Aurelia Modrego from Carlos III University. I owe you a lot. Our days in Getafe and Colmenarejo are moments in my life where I have learnt and enjoyed enormously. In the group of Aurelia I was lucky to meet and work with Emma, Myrna and Myriam. With them I learnt the importance of teamwork, respect and understanding. Thank friends for always being there.

Everyone I've met during these years in Leiden has contributed greatly to the daily life of the thesis: the pleasant and stimulating dinners on Saturday evenings with Victor, Beatrice and Judith; the parties with the Tijuana group; the evenings with Elisabetta; the Tuesday lessons with Annelies; the intensive yoga courses with Isabel; Ulrike and Yvonne, my German connections; Maria Garcia and the dinners with her; my long walks with Leticia. I thank you all for being there.

Veronica and Bruno, It was a wonderful coincidence that we met. And with Veronica and Bruno arrived Elena, Giulio, Maria, César and Alicia. We have spent so many moments together during these years that I cannot even decide which one I should mention here. Each of you has been great support, each of you are part of this thesis. It is a privilege to have you as friends. Thanks pandi.

Some of my friends in Spain have been so present these years that sometimes I felt the physical distance did not exist. Thanks Lorena, Alvaro, Patricia, and Juan. And especially thanks to Myriam, my other sister, for the long road we have walked together.

Gracias abuelos por conectarme con mis raíces y darme fuerza. Gracias a Mari Cruz y Sengan, siempre os llevo conmigo. Gracias Maria José, Soledad, Elena, Cristina, Lourdes, Anselmo, y Paco. Desde el Sur siempre me habéis mandado calidez y cariño.

Y claro, sin mis padres y mi hermana nada hubiera sido posible. Gracias padre, gracias madre, por todo vuestro esfuerzo y sacrificio durante muchos años. Vuestro amor por la enseñanza y el conocimiento son parte de esta tesis. Pilar ni si quiera puedo agradecer con palabras todo lo que siempre has hecho por mi. Hermana es un honor tenerte como compañera en este viaje de la vida. Lola se incorporó a la tesis en los últimos tiempos, pero ¡qué gran ayuda! Gracias sobrina. Gracias Andreas por darme caña y recordarme cada dos por tres que debía terminar la tesis.

Last, but not the least, I would like to thank Moos. You took my hand and walk with me the moment I really need it. I learn so many things with you that it would be impossible to summarize in a thesis. Moos heel erg bedankt. Je bent de beste leraar die ik ooit gehad heb.

Seguimos caminando….

Table of Contents

1 General Introduction

1.1 Introduction

Knowledge has always been at the core of economic growth and social welfare. The capacity to invent and innovate, to create new knowledge and new ideas that later become part of products, processes and organizations, has always fostered development. Many organizations and institutions have been effective in creating and disseminating knowledge: from the corporations of the Middle Ages to the large firms at the beginning of twentieth century, and from the Cistercian abbeys to the royal academies of science that began to appear in the seventeenth century (David & Foray, 2003) .

But even though knowledge has always been important for economic development, the term "knowledge-based economy" is quite recent (OECD, 1996), and thus marks a break and introduces a discontinuity with respect to previous periods. Historical explanations of the abundance (or scarcity) of natural resources have lost much of their effectiveness in explaining disparities in productivity and growth across countries. In contrast, the improved quality of physical equipment and human capital represents a better explanation, as this relates to the creation of “new knowledge and new ideas and incorporate them into the equipment and people” (David & Foray, 2003). Since the beginning of the twentieth century a new characteristic of economic growth has been detected which consists in the growth of the share of intangible capital as compared to tangible capital (Abramovitz and David, 1996). Part of the intangible capital consists of investments in training, education, R&D activities, information and coordination; this means investments devoted to the production of knowledge and human capital.

The knowledge-based economy arises when a group of people produce and exchange new knowledge intensively with the help of information and communication technologies. Therefore, three elements may be distinguished: (1) the production and reproduction of new knowledge is taken up by a significant number of community members; (2) the community creates a "public" space-sharing knowledge movement through new information technologies; and (3) the communication to encode and transmit the new knowledge is intensive.

One of the main issues in a knowledge-based economy is to measure effectiveness in the production, measurement and use of knowledge. Therefore, in this context it is not the mere accumulation of knowledge that is important but the ability to use it in meaningful ways (OECD, 1996). However, knowledge is a concept that is difficult to quantify and/or put a price on. Traditionally, knowledge has been classified as

basic or *applicable*, and depending on how it is stored knowledge can also be classified as *codified* or *tacit*. According to van Raan's (2004) definition, codified knowledge is 'archived & publicly accessible', and the non-codified or tacit knowledge is 'craftsmanship'. Both codified and non-codified knowledge are essential parts of the knowledge-generating processes: codified knowledge helps diffusion and exchange, and non-codified knowledge, located in individuals, is essential to the understanding and use of the former kind.

In knowledge-based economies the science system increases in importance. Public research laboratories and universities are at the core of the science systems, a core extended to government science institutions and research councils, R&D intensive companies, and the supporting infrastructure (OECD, 1996). Consequently, the professional communities that are most engaged in the knowledge-based economy are scientific communities. These are indeed the communities in which, by definition, most members are producers of knowledge to be shared (Dasgupta and David, 1994) and that historically have always been pioneers in the use of new information technologies. Scientific production, however, embedded in the knowledge process, has a complex structure, shaped by technical and social influences (Schmoch, Schubert, Jansen, Heidler and von Gortz, 2010). This complexity is related to the trend towards multidisciplinary and interdisciplinary research, and the increased desire for and necessity of collaboration between researchers. (PREST, 2000). Thus, research is a collective effort combining diverse actors, competences and capabilities, and emphasizing the collective setting, the interface between individual researchers and research institutions (Laredo, 2003).

One of the major forms of scientific output, embedded in this complex system of scientific production, is scientific publishing. Scientific publications represent a specific but immense collection of codified knowledge that can be easily disseminated and absorbed among knowledge users. They can also be easily stored for future use. Publications also provide an important indication of what is leading-edge research, and where it is being performed (Hauser & Katz, 1998). But scientific publications themselves are also an excellent platform for studying how knowledge is shared and disseminated inside the scientific community.

The interconnections between scientific publications (e.g. citations given and received from one paper to another) and inside them (e.g. researchers co-authoring papers) allow us to study the way in which scientists create and share new knowledge by means of network analysis, which may help

to reveal the conditions behind the successful share and transfer of knowledge. Derek de Solla Price (1965) already showed the structure of science as a network of interconnected publications. In the last few decades a diverse group of scientists, including mathematicians, physicists, computer scientists, sociologists, biologists as well as information scientists, have been actively working on network theory in an effort to understand and explain its properties. In network theory terminology the number of citations of a paper is the *in-degree* of a paper, being a local property of the citation network. This quantity gives information about the characteristics of the network around the nodes, but it does not help to uncover the highly clustered structure of the scientific network. In order to understand the complexity behind knowledge production we also need to study the structure of interconnected publications, otherwise we may in fact be missing some important and crucial phenomena. Traditionally, the first approach to analyze the structure underlying a network is to make picture of it. During the last years there has been a rapid development in the field of information science applying different techniques to visualize bibliometric networks. Next to visualization techniques ('mapping') the structural characteristics of scientific networks can be studied using measures and metrics developed in network theory through the years. The developments in the last years in network theory is helping to incorporate these measures to the studies of scientific networks with the goal of getting to understand better the process of knowledge creation and sharing.

This thesis originated from the need to identify groups of related nodes within the collaboration and citation networks. In the study of collaboration networks the main goal is to identify research groups, potential research groups or patterns of collaboration. The analysis of citations networks through specific measures and metrics, on the other hand, makes it possible to identify main lines of research through the years. Thus, such analyses improve our understanding of the growth and decline of fields, including phenomena such as paradigm shifts and emerging research themes. Network measures and metrics also allow for the identification of important nodes (e.g., journals, articles) embedded in the citation net.

1.2 Network analysis

The aim of this chapter is to show the main lines of the historical developments in network theory, together with a number of representative concepts. The objective is to get a feeling for the kind of properties of networked system that can be measured or modeled and

how these properties are related to practical issues presented in this thesis.

Historical developments

In 1736 the mathematician Leonhard Euler took an interest in a mathematical puzzle inspired by an actual situation called the Seven Bridges of Königsberg. The city of Königsberg, Prussia (now Kaliningrad, Russia) on the Pregel River included two large islands which were connected to each other and to the mainland by seven bridges (Figure 1). The popular question at that time was whether it was possible to walk a route that crossed each bridge exactly once, and then returned to the starting point. Euler proved with a graph that this was not possible. A graph is a mathematical object consisting of points (nodes, vertices) connected by lines (links, edges, arcs). In Euler's graph the four nodes representing the four pieces of land were connected by seven edges representing the seven bridges (Figure 2). As Newman, Barabasi and Watts (2006) explain, the bridge problem can be phrased in mathematical language as the question of whether there exists an Eulerian path in the network. An Eulerian path is a path that traverses each edge exactly once. Euler's proof is considered by many to be the first theorem in a mathematical field called graph theory, which is the main mathematical framework in which properties of networks are described (Harary, 1996).

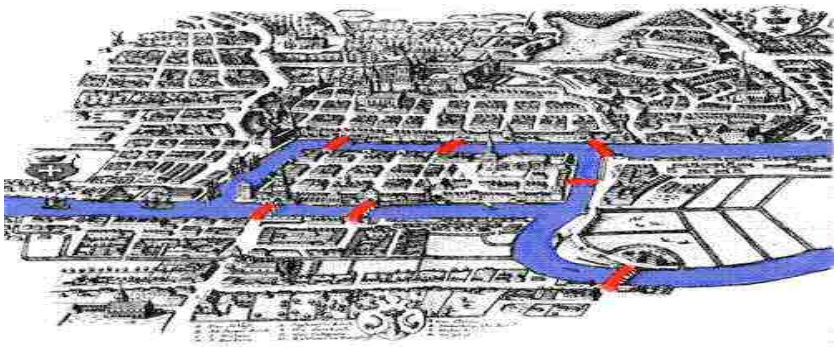

Figure 1. Map of Königsberg in Euler's time showing the layout of the seven bridges, highlighting the river and the bridges

The strength of a graph is that the nodes and the edges can be almost anything, since many systems can be simplified to a network structure while maintaining complexity (Rosvall, 2006). The complexity can be retained because a complex system is made up of a large number of

components, or agents, interacting in such a way that their collective behavior is not a simple combination of their individual behavior (Newman, 2002). However, it is important to remark that to be able to abstract a system into its underlying network the units have to be unique, such as for instance humans, proteins, scientific publications, or web pages. A system containing interchangeable units, such as atoms or electrons, cannot be reduced to a network. By abstracting away the particulars of a problem, network theory is capable of describing major topological features with a clarity that would be impossible if all the details were retained. This is why network theory has expanded outside its original domain of pure mathematics (Newman, Barabási, & Watts, 2006).

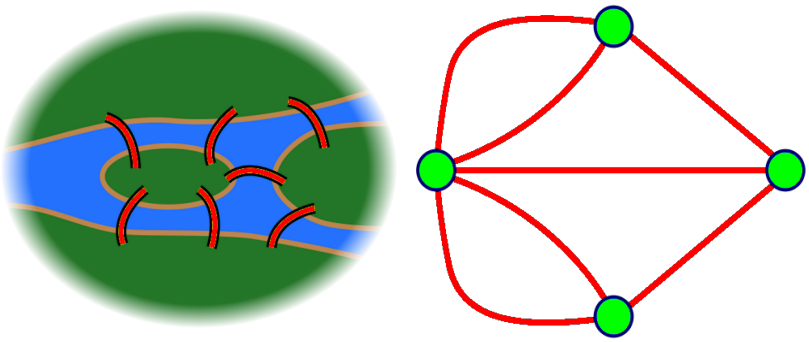

Figure 2. Left: A simplified representation of the pattern of the river and bridges in the Königsberg bridge problem. Right: the corresponding network of vertices and edges. (Source: Newman, Barabási, and Watts, 2006)

From the 1930s the mathematical language of graph theory has been adopted by social scientists to help them to understand data from ethnographic studies (Borgatti, Mehra, Brass, & Lambiaca, 2009). Since then, social network analysis has emerged as an integrated scientific speciality concerned with the structural analysis of social interaction (Hummon & Carley, 1993). Social network analysis concentrates on the interpretation of the social nature of the nodes, and on the edges between them (Marion, Garfield, Hargens, Lievrouw, White, & Wilson, 2003).

At the beginning of the 1950s mathematicians began to think of graphs as the tool to study the spread of various 'modes of influence' – especially information and diseases. The structural properties of networks, particularly their connectedness, became linked with behavioral characteristics such as the expected size of an epidemic or the possibility of a global information transmission. This line of research also included

the notion that graphs should be regarded as stochastic rather than purely deterministic objects, so that graph properties can be thought of in terms of probability distributions – which is the link with the new developments in network theory in recent years (Newman, Barabási, & Watts, 2006).

The novelty of recent developments in network theory is that researchers, mainly physicists, have started to use the principles of statistical mechanics to analyze large networked structures (Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002; Newman, Barabási, & Watts, 2006). This 'complex network theory' mainly concentrates on analyzing degree distributions, clustering coefficients, and theoretical mathematical models to explain empirical findings. There have been many discoveries and developments in network theory during the last decade driven by the ever increasing availability of empirical data. Probably the most surprising finding is that many real networks, independent of their age, scope, and function, converge to structures with similar properties (Barabási, 2009).

Basics of Network Theory

Often a first step in analyzing the structure of a network is to make a picture of it. A network – 'graph' in mathematics- is made up of points, called *nodes* or *vertices*, and lines connecting them, usually called edges. Figure 3 shows the example of a network.

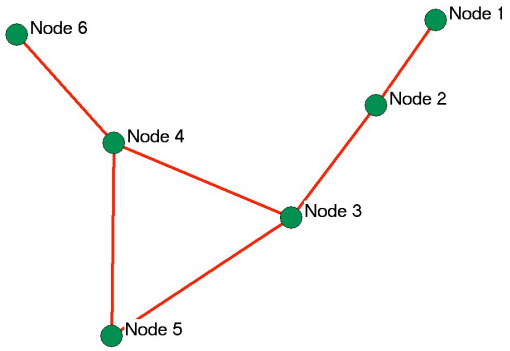

Figure 3. Example of a network

The structure of a network is described by its adjacency matrix ***A***, which in the simplest case is a *n x n* symmetric matrix, where *n* is the number of nodes in the network and A_{ij} are the elements.

$$A_{ij} = \begin{cases} 1 \text{ if there is an edge between vertices } i \text{ and } j, \\ \\ 0 \text{ otherwise.} \end{cases}$$

$\boldsymbol{A}$ is the adjacency matrix of the network in Figure 3.

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

This is the case of a symmetric matrix, since there is an edge between i and j and an edge between j and i, thus $A_{ij} = A_{ji.}$ The symmetric matrix represents an *undirected* network. If the edges have directions, i.e., if the edge points from i to j or from j to i (but not both) the network is directed. In the case of directed networks $A_{ij} \neq A_{ji}$.

The edges can also be weighted to represent stronger and weaker connections. Thus, the adjacency matrix can be generalized to values other than unity to represent the strength of the connections.

Network measures

An example of an important class of network measures is centrality. Centrality is a family of node level properties relating to the structural importance or prominence of a node in a network (Borgatti, Mehra, Brass, & Lambiaca, 2009). Social network analysts have studied and developed measures of centrality for a long time; as a result there is a wide range of concepts and definitions about what it means to be central to a network.

The simplest centrality measure is the degree centrality. The degree of a node in a network is the number of edges attached to it. For instance, in a friendship network between individuals the degree of a person will be the number of friends the individual has within the network.

If the edges are directed a node will have two degree measures: in-degree and out-degree. A well-known measure in a citation network is the in-degree of a paper: the number of its (received) citations. It is the basic standard measure for quantifying impact (De Solla Price, 1965; Egghe & Rousseau, 1990; Garfield, 1972; Lambiotte & Panzarasa, 2009; Wuchty, Jones, & Uzzi, 2007). On the other side the out-degree of a paper is the number of references given by the paper. A high out-degree is found, for instance, in review papers.

The corresponding global description of the network as a whole is the degree distribution, where the tail of the distribution follows power-law function in the case of citation networks (De Solla Price, 1965; Redner, 2005; Clauset, Shalizi & Newman, 2009) and co-authorship networks (Newman, 2001). Actually, many networks contain a small but important number of nodes of an unusually high degree. The effects of these nodes on the performance and behavior of network systems is one of the main lines of research nowadays, since information on these effects can help to avoid the expansion of a disease or of a targeted attack on internet (Newman, 2010).

There is a second version of the degree measure, which is more complicated but based on the same idea, called 'eigenvector centrality'. This eigenvector centrality acknowledges that not all edges are equally important, and provides each node a centrality depending on the number and the 'quality' (weight) of the connections. The eigenvector centrality turns out to be a useful measure in many situations. Actually, already in the 1970s, Pinski & Narin (1976) and Geller (1978) developed a measure of journal impact based on the eigenvector centrality. Their algorithm considered not only the number of citations from one journal to another, but also the prestige of the citing journal based on the average journal impact Journals that receive many citations from other prestigious (i.e., highly cited) journals are considered highly prestigious themselves. By iteratively passing prestige from one journal to the other a stable solution is reached which reflects the relative prestige of journals (Bollen, 2006). This way of measuring prestige is also behind the recent PageRank algorithms used to evaluate the status of web pages. The PageRank is calculated by an iterative algorithm which analyzed prestige values from one web page to another and converges to a stable solution (Brin & Page, 1998; Page, Brin, Motwani, & Winograd., 1998; Pillai, Suel, & Cha, 2005). Kleinberg (1999) also worked on an algorithm to increase the effectiveness of web search engines, using the concepts of hubs and authorities. Hubs & authorities are formal notions of structural prominence of vertices in directed graphs (Brandes & Willhalm, 2002).

We will come back to the characteristics of hubs and authorities in next section.

Another way of measuring the central position of a node in a network is called betweenness. A node with a high betweenness centrality is a node that appears very often in the shortest path that connects any two other nodes from the network. Freeman in 1977 developed this centrality measure and is considered as a measure of the influence of a node in the network in terms of information flow.

Another interesting and well-known concept in network theory is called the small-world effect. The geodesic distance between two nodes in a network is defined as the minimum number of nodes (shortest path) one has to pass through to get from one node to another. The small-world effect shows that in most networks the mean geodesic distance between node pairs is surprisingly short compared to the size of the network as a whole. The idea was first explored mathematically by Pool and Kochen during the 1950s (Pool & Kochen, 1978), by Milgram during the 60s (Milgram, 1967), and by Watts and Strogatz during the 90s (Watts & Strogatz, 1998). The mean geodesic distance varies with the type of network, but the basic principle that you can go from an arbitrarily chosen node to any other node in just a small number of steps is well documented in a wide array of systems (Newman, 2008). The small-world effect has important consequences, for instance for the Internet. One of the reasons why Internet functions is because any computer in the network communicates by only a few "hops" over optical or electronic data lines. In practice, data packets sent over the Internet travel typically in the range of about ten to twenty hops long. The performance of the Internet network would be terrible if the packets had to make a thousand hops instead (Newman, 2010).

Another important network concept is the clustering or network transitivity (Watts & Strogatz, 1998; Watts, 1999). A network shows clustering if the probability of two nodes being connected by an edge is higher when these nodes have a common neighbor. Eckmann & Moses (2002) showed there is a close relation between highly clustered regions of a network and the existence of communities. The way a network breaks down into communities can reveal levels and concepts of organization that are not easy to see without network data, and it can help us to understand how a system is structured (Newman, 2010). The development of methods for finding communities within networks is a prosperous sub-area of the network field, with a large number of different techniques under development. However, methods for understanding

what these identified communities really mean are still in the very early stages of development (Newman, 2008).

1.3 Bibliometric analyses

Nowadays knowledge producers, especially the public research laboratories and universities, have to deal with different and sometimes contradictory demands from society. They have to face unpredicted policies in education and research mainly linked to budget reductions, as well as an accelerating rate of knowledge growth together with the internationalization of the knowledge process itself. The picture becomes even more complicated if we consider that research itself is a complex and collective effort combining various actors, competences, and capabilities. It is essential that academics, research managers, and policymakers stay abreast of the way research works and the impact that science policy and research management have on research. This is the reason why methods for the study of research performance – including bibliometric analyses – should be conceived as an interdisciplinary effort, aimed at integrating perspectives, insights, and findings from a series of relevant scientific-scholarly disciplines.

Bibliometrics, the quantitative analysis of bibliographic data, plays an important role in the study of research performance. The experience gained by bibliometricians in the analysis of scientific publications, and the criteria for their usefulness expressed by research management and policy makers, keeps the bibliometric analyses in the realms of both theoretical reflections as well as empirical research of an application-oriented nature. The vast information contained in scientific publications, the different analyzing techniques available, and the different questions we want to help answer, require detailed analysis of scientific communication.

For the purpose of the work presented here we divided the bibliometric studies and analyses carried out at the Centre for Science and Technology Studies (CWTS) into three research lines that are interconnected and complement each other: (1) performance analysis based on direct counts of citations received by publications; (2) bibliometric mapping of science; and (3) detailed collaboration and citation analysis using the network of linkages between publications. Two of these lines have been at the core of the CWTS research and studies for decades now: performance analyses primarily based on bibliometric indicators as part of processes for the assessment of research performance, and bibliometric

mapping of science to unravel the difficult-to-classify science system and to support the assessment of research performance.

The third line, detailed collaboration and citation analysis, is the most recent. This approach originated from a need to identify groups of related nodes inside the collaboration and citation networks. Regarding detailed collaboration analysis, the main goal is to identify research groups, potential research groups or patterns of collaboration. The detailed citations analysis, on the other hand, makes it possible to identify main lines of research through the years and thus improves our understanding of the growth and decline of specific fields, including phenomena such as paradigm shifts and emerging research themes. The detailed citation analysis also allows for the identification of important nodes (e.g., journals, articles) embedded in the network.

Performance analysis

Performance analysis, based on publication output and citations received, is used to assess the performance of research communities. The process of citation is a complex one, and certainly does not provide an "ideal" monitor on scientific performance. This is particularly the case for a statistically low aggregation level, for instance, an individual researcher. But the application of citation analysis to the work, the "oeuvre", of *a group as a whole over a longer period of time*, does yield in many situations a strong indicator of scientific performance, and in particular of scientific quality. An important and absolutely necessary condition is that applied citation analysis is part of an advanced, technically highly developed bibliometric method. Bibliometric indicators are used to assess the research output of countries, universities or research institutions, and departments or research groups (Moed, De Bruin & Van Leeuwen, 1995). The work done by Garfield (1979), Martin & Irvine (1983), Narin (1990), Van Raan (1997), and Schubert, Glänzel & Braun (1989) shows the importance and strength of the performance indicators when it comes to assessing the output of a research unit.

Performance has three central aspects: activity, productivity, and impact. Connecting the scientific output of a research unit to the number of citations received (in-degree in the citation network) provides us with an indicator of impact, influence, or at least visibility (Noyons, 1999). CWTS has been working for many years on improving the bibliometric indicators and adapting them to the specific demands of researchers, research managers, and policy makers (van Leeuwen, 2004). Many studies support the use of the bibliometric methodology developed at CWTS for assessing performance in different fields such as physics

(Rinia et al., 2001), biology (Nederhof & Visser, 2004), electrical and electronic engineering (Van Leeuwen et al., 2000), chemistry (Van Leeuwen et al., 2003), humanities (Nederhof, 2006; Tijssen et al., 2006), medicine (Tijssen et al., 2002), and social and behavioural sciences (Nederhof, 2006).

Bibliometric mapping of science

Each year about a million scientific articles are published. How to keep track of all these developments? Are there specific patterns 'hidden' in this mass of published knowledge, at a 'meta-level', and if so, how can these patterns be interpreted (Van Raan & Noyons, 2002)? Structuring science is about identifying fields, sub-fields, and research themes and relating them to each other. The mapping of science by means of co-word and co-citation approaches has also been part of bibliometric studies for a long time (Braam, Moed & van Raan, 1991a, 1991b; Callon, Law, & Rip, 1986; Chen, 2003; Garfield, Pudovkin, & Istomin, 2003; Small, 1999; Tijssen & van Raan, 1989). This became necessary because the traditional science classification system is imperfect, especially for highly multidisciplinary environments, and it helps to assess performance.

The data behind science mapping are bibliometric networks and until now this technique has been used mainly with co-occurrence networks (based on keywords in publications) and with co-citation networks (based on citations received and given by publications, authors or journals). Because these maps usually cover many publications, a simple network representation, i.e. a set of nodes and edges, is of no use since the human eye can not catch the information in a big and dense network graph (Newman, 2010). This is why more advanced visualization techniques that allow the representation of the network data in comprehensible maps are used. At CWTS the work carried out through the years (i.e., Noyons, 1999; van Eck & Waltman, 2007, van Eck & Waltman, 2010) shows the importance of this procedure as a research management and science policy tool.

Collaboration and Citation Analyses

- *Detailed Collaboration Analysis*

It is often said that in recent decades there has been a sharp increase in the number of papers that involve collaboration among researchers detrimental to papers without collaboration (Hicks and Katz, 1996). Part of the reason for this increase in the proportion of collaborative work, lies in the need for more specialized and concentrated resources, together with an increase in interdisciplinarity (Gibbons et al., 1994). Moreover,

numerous studies have highlighted the positive relationship between research productivity and quality on the one hand, and collaboration between many researchers on the other (e.g., Lawani, 1986, Peters & van Raan, 1994). In addition, the characteristics that influence the intensity of collaboration are various, depending, for instance, on scientific discipline, institutional level, or geographic level (local, national or international) (Katz and Martin, 1997).

Bibliometric analyses play an important role in measuring these tendencies, and a number of studies have been carried out since the 1990s. The co-authorship data were used in many studies to measure collaboration (e.g., Persson & Beckmann, 1995; Martin-Sempere et al., 2002; Melin & Persson, 1996; Bordons & Gomez, 2000; Van Raan, 1998; Seglen & Aksness, 2000). From around 2000 several researchers began the construction of large-scale networks using co-authorship data in mathematics (Barabási et al., 2002); biology, physics and computer science (Newman, 2001); and neuroscience (Barabási et al., 2002). During the last decade network researchers have been working to reveal the highly clustered nature of scientific production, showing that co-authorships networks are made up of several dense groups of nodes, called 'communities' (Lambiotte & Panzara, 2009).

Our daily work with research managers in highly interdisciplinary research centers shows the need for new approaches to help them reorganize their centers, which often are still organized in traditional, disciplinary ways. The novelty of our approach is that we have combined different methods in order to identify communities and functional or potential research groups. Regarding collaboration analysis we used two techniques developed in network theory:

- A technique to identify 'regions' between the nodes, called k-core. A k-core is a subgraph in which each node is connected to at least a minimum fixed number (k) of the other nodes in the subgraph (Seiman, 1983). The k-core approach allows actors to join the group if they are connected to k members, regardless of how many other members they may not be connected to (Wasserman & Faust, 1994).

- The Girvan and Newman algorithm to identify the communities and groups based on co-publication networks (Girvan & Newman, 2002; Newman, 2004; Newman & Girvan, 2004). This Girvan-Newman uses the edge betweenness measure as the basis of their algorithm. Based of the same idea of the node betweenness developed by Freeman (see section 1.2), the edge betweenness of an edge measures the times an edge is used in the shortest paths that connect two other

nodes from the network. The edges that connect highly clustered communities have a higher betweenness so cutting these edges should separate communities. The method finds divisions of networks into closely knit groups by looking for the edges that connect groups (Lusseau & Newman, 2004).

The work presented here shows how we can identify communities and groups ('functional research groups', see Seglen & Aksnes, 2000; Calero et al., 2006) by using network analysis techniques to analyze collaboration data and combine the results with other bibliometric techniques (bibliometric mapping of science and performance analysis) This approach may lead to a better understanding of how complex interdisciplinary organizations work and may therefore support research managers to reorganize their organization in a more efficient and practical way.

- *Detailed Citation Analysis*

Citation network analysis began with the study by Garfield, Sher & Torpie (1964) of Asimov's history of DNA. Isaac Asimov described in his book "The Genetic Code" the major scientific developments that enabled the duplication in a laboratory of the protein synthesis process under control of DNA. Garfield and colleagues created a citation network taking as starting point the papers where the main milestones mentioned by Asimov where published and the citations between these papers as links. They showed that there was "a high degree of coincidence between an historian's account of events and the citation relationship between these events". In terms of citations, the representations of fields or areas of specialization are not just 'formless' sets of articles. On the contrary, they represent sets of papers with a particular structure that emerges from the citation practices of the researchers active in that field. They emphasize the importance and visibility of certain theoretical and methodological approaches while marginalizing others. We could say that citation practices represent a "knowledge-construction" process that outlines the manner in which we think about and engage with our research.

In all scientific fields there are key concepts that form the basis for theoretical developments through the years. Researchers from the same specialty tend to cite each other in order to position their work in the field on the basis of previous knowledge. Scientific knowledge is assumed to increase over time following a "smooth path"; the papers that introduce important new insights are cited until they are modified or contradicted by new results. The scientific revolutions, i.e., sudden paradigmatic

changes resulting from new insights (Kuhn, 1969), are reflected by abrupt changes in the citation network. In this context, following Small (1978), a cited document stands for a concept. Highly cited documents have a significant content that is shared by a community of scientists.

The citation network has enabled us to analyze the data from two perspectives in terms of time: longitudinal studies and cross-sectional studies. The techniques of longitudinal network analysis show the changes over time in the connectedness of the system. In the evolution of knowledge, phases of consolidation of past results coexist with the exploration of new approaches. One of the techniques is called main path analysis. The second technique is a cross-sectional analysis of the citation network at a well-defined time, using a specific algorithm to identify prominent nodes in the citation network called hyperlink-induced topic search (HITS). These two perspectives are important because they highlight different parts of a citation network.

Main Path

The main path analysis makes it possible to unravel the dynamics of convergence and divergence between 'investigation streams' (Ramlogan et al., 2007). If knowledge flows through citations, a citation that is a necessary step in many paths between many articles is more important than a citation that hardly plays any role in linking articles (De Nooy, Mrvar & Batagelj, 2005). Among all possible "chains" of citations, from the most recent to the oldest, the network algorithm computes the paths that are most frequently encountered. These paths can be regarded as the backbones of a research tradition (Hummon & Doreian, 1989, 1990; Hummon & Carley, 1993; Batagelj, 2003; De Nooy, Mrvar & Batagelj, 2005). These results identify the path that is most frequently used to 'walk' from the present to the past (back in time) in a 'field' of papers; this path is called the 'main path'. It is important to stress that this method does not involve the absolute count of maximum numbers of citations received, but the simultaneous computations of all possible paths through the whole dataset and the choice of the one that is most frequently used through time (Mina et al., 2007).

HITS - Hubs and authorities

The concept at the basis of 'hubs' and 'authorities' in a network can be dated back to Pinski and Narin (1976). They proposed to measure the prominence of scientific journals by taking into account not simply the number of citations that a journal receives, but also the prestige (in terms of citations received) of the journals that cite it. Journals that receive many citations from prestigious journals are considered highly

prestigious themselves and, by iteratively passing prestige from one journal to another, a stable solution is reached which reflects the relative prestige of journals (Bollen et al, 2006). This way of measuring prestige is the basis of the algorithms for evaluating the status of web pages developed by Brin and Page (1998) and Kleinberg (1999). Kleinberg (1999) constructed a 'centrality' algorithm to increase the effectiveness of web search engines.. This algorithm, called hyperlink-induced topic search (HITS), is based on the idea that there are two types of important nodes in a directed network: hubs and authorities. Hubs and authorities are formal notions of structural prominence of vertices in directed graphs (Brandes & Willhalm, 2002). The algorithm gives each node in a network an authority centrality and a hub centrality. A hub is a node with a large number of links (hub centrality). A node with high authority centrality is one that has many links with hubs, i.e., many other vertices with high hub centrality. The characteristic of a node with high hub centrality is that it points to many nodes with high authority centrality (Newman, 2010).

Authorities are nodes that contain useful information on a topic of interest; hubs are nodes that tell us where the best authorities are to be found. An important scientific paper (in the authority sense) is one that is cited by many important reviews (in the hub sense). On the other hand, an important review is one that cites many important papers. However, "ordinary" papers can also have high hub centrality if they cite many other important papers, and papers can have both high authority and high hub centrality. The reviews, too, may be cited by other hubs and hence have high authority centrality as well as high hub centrality (Newman, 2010).

As an example, Figure 4 below shows the citation network between 10 papers labeled by the publication year (ten different years). The lines (directed edges) show the citation relation between the papers. The direction of the arrow indicates if a paper is cited by (receiving an arrow) the paper on the other side of the line. Notices how the citation flow is related with the year of the publication, this means that for instance the paper from 2005 can not be cited by the paper from 2001. From the citation network the nodes with two highest authority centrality measures and the two highest hub centrality measures have been highlighted. The 2002 paper (blue square) is one of the two papers with the highest authority centrality while the paper from 2008 is one of the two papers with the highest hub centrality measure. The paper from 2005 (the yellow diamond) is the paper that has the other highest hub centrality measure and the other highest authority measure. As the graph shows the paper from 2005 is not only citing to many papers but it is citing also to the one from 2002 that it is an important paper (in the authority sense), and at the

same time the paper from 2005 is not only cited by many papers but one of them (2008) is an important paper (in the hub sense).

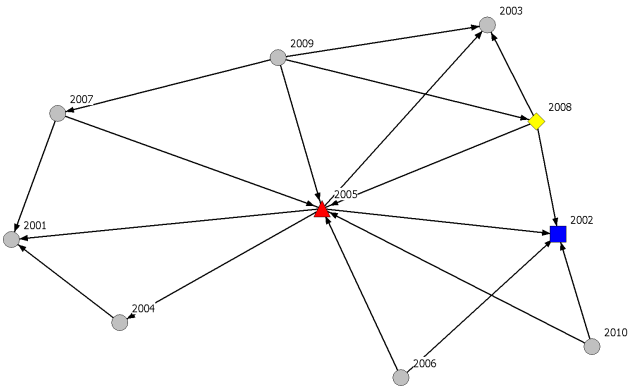

Figure 4. Citation Network between 10 papers

For the software Pajek, Batagelj and Mrvar adapted Kleinberg's hubs/authorities algorithm. The results from the analyses presented in this thesis are based on Pajek.

1.4 Linkages in bibliometric studies and thesis outline

The work presented in the next chapters shows how in the past years, due to the increasing need to understand the collaboration and citation process and the developments in network theory, we have developed a third line of research embedded in the bibliometric analyses and combined with the other two lines. In Figure 5 we illustrate this combination of the three lines of bibliometric analysis.

In general terms we could say that performance analysis with bibliometric indicators yields *specific output- and impact-related information* about a specific entity (e.g., country, university, department), whereas science mapping yields *more general structural information* about a research field. Our analyses allow us to link the two approaches. As Noyons, Luwel & Moed (1999) have shown, the mapping procedure can improve the performance analyses so that the performance in a research field may be investigated in more detail. For instance, the position of a research institute on the map. On the other hand, the performance indicators contribute to the validation of the structures

obtained by means of the science maps. Still, there was a gap between the specific and the general analysis. A gap an "in-between" of what happens in the networked system in terms of levels, concepts and 'natural' communities within organizations that are not easy to see without analyzing publications using network measures and metrics to understand how the scientific system is structured. The above also explains the difference in bibliometric practice between the more general structural analysis by science mapping, and the more detailed structural analysis by the network approaches.

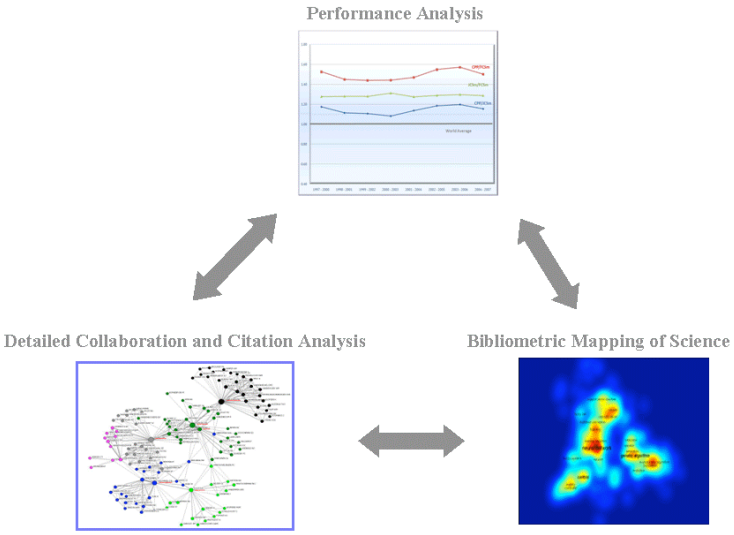

Figure 5. The three main CWTS bibliometric analyses

The research described in this thesis aims to establish the use of detailed collaboration and citation analysis combined with other forms of bibliometric analysis as a tool enabling a better understanding of the organization of scientific communities and the way knowledge is spread inside scientific communities. In this perspective there are three key questions that we address in this thesis:

Can we identify communities, research groups and potential research groups?

The answer to this question is crucial for helping research managers and policymakers to organize complex organizations in a more efficient and practical way.

Can we identify main lines of research through the years, and the articles that linked them into a research tradition that can be considered as the backbone of the field?

The answer to this question depends on a better understanding of the growth and decline of specific fields, including phenomena such as paradigm shifts and emerging research themes.

Can we identify important nodes that play a key role in the citation networks?

The identification of important nodes (e.g., journals, articles) embedded in the network is related to understanding how information flows.

In **Chapter 2** and **3** we present two approaches to identify research groups in a particular research field, or inside an organization. Both approaches deal with the complex issue of the position of research groups within a changing structure of scientific research. In particular, in Chapter 2 we identify and classify clusters of authors to represent research groups by means of a combination of bibliometric science mapping techniques and detailed network-based collaboration analysis. We present two types of outcomes: actual research groups and potential research groups. The former enable us to define research groups beyond the formal organizational structure, and the latter can be used to identify potential partners for collaboration.

In **Chapter 3** we combine data on bibliometric indicators with detailed collaboration analysis to examine the formal organization of a University Hospital. Allowing the co-publication network itself to identify communities and groups inside this interdisciplinary research centre (in fact a kind of self-organization) may lead to a better understanding of how this complex organization works and how to reorganize research in a more efficient and practical way.

In **Chapter 4** we present a study on research cooperation within multinational enterprises (MNE) in the bio-pharmaceutical industry. We use the publications of the MNEs to examine structural factors characterizing research cooperation networks within the industry at the level of major geographical regions (North America, Europe, Pacific-

Asia), with a breakdown into within-MNE and between-MNE network linkages.

In **Chapter 5** we present bibliometric characteristics of the world and European universities with the largest scientific output in terms of publications. We compare US universities with European institutions for a number of different aspects, for instance countries with a strong concentration of academic research activities within a core group of universities and countries with a more even distribution of research over institutions. We present a ranking of universities based on indicators calculated for *all* research fields with a ranking for just one specific field (Oncology). Here we distinguish between general, broad universities and specialized universities. We also present results for rankings based on a single indicator with collaboration maps combining network analysis and a series of indicators.

In **Chapter 6** we present a study in which we combine bibliometric science mapping based on co-word network analysis and a specific analysis inside the citation network to investigate the process of knowledge creation and dissemination through scientific publications. We analyze the citations of a very influential paper that introduced a term in a field to identify the articles that influenced the research for some time and to link them to a research tradition that can be considered the 'backbone' of the field.

In **Chapter 7** we present a method based on the application of network theory to citation networks in order to identify the most important journals related to a given journal, the 'seed journal'. In just one simple network map we can sketch the relevant citation environment of these seed journal. This approach is of interest to publishers, librarians, scientists, and to science policy makers.

Finally, in **Chapter 8** we summarize our conclusions and illustrate the prospects for future research.

References

Abramovitz , M., and David , P.A. (1996). Technological change and the rise of intangible investments : the US Economy's growth-path in the twentieth century, in D. Foray and B. A. Lundvall (eds.). Employment and Growth in the Knowledge-based Economy, OECD, París : OECD.

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1), 47-97.

Barabási, A.-L., Jeong, H., Ravasz, E., Néda, Z., Schuberts, A., and Vicsek T. (2002). Evolution of the social network of scientific collaborations, Physica, A 311: 590–614.

Barabási, A.-L. (2009). Scale-Free Networks: A Decade and Beyond. Science, 325, 412-413.

Barjak, F., and Robinson, S. (2007). International collaboration, mobility and team diversity in the life sciences: Impact on research performance. Social Geography Discussion, 3, 121–157.

Batagelj, V. (2003). Efficient Algorithms for Citation Network Analysis. Preprint Series. Univ. Ljubljana, Inst. Math., 41 (897),1-29.

Batagelj, V., and Mrvar, A. Pajek - Program for Large Network Analysis. Home page http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Borgatti, S.P., Mehra, A., Brass, D.J., and Labianca, G. (2009). Network analysis in the social sciences. Science, 323, 892-895.

Bollen, J., Rodriguez, M.A., and Van de Sompel, H. (2006). Journal status. Scientometrics, 69 (3), 669-687.

Bordons, M., and Gómez I. (2000). Collaboration networks. in science. In: H. B. Atkins, B. Cronin (Eds.), The Web of Knowledge: A Festschrift in Honor of Eugene Garfield, Information Today, Medford, NJ, 197-213.

Brandes, U., and Willhalm, T. (2002). Visualization of bibliographic networks with a reshaped landscape metaphor. Joint Eurographics-IEEE TCVG Symposium on Visualization, D. Ebert, P. Brunet, I. Navazo (Editors). http://algo.fmi.uni-passau.de/~brandes/publications/bw-vbnr1-02.pdf .

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30, 107-117.

Calero, C., Buter, R., Cabello, C., and Noyons E. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. Scientometrics, 66(2), 365–376.

Chen, C.(1999). Visualising semantic spaces and author co-citation networks in digital libraries. Information Processing and Management, 35, 401-420.

De Nooy, W., Mrvar, A. and Batagelj, V. (2005). Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press.

De Solla Price, D.J. (1965). Networks of scientific papers. Science, 149(3683), 510-515.

Dasgupta, P., and David, P.A. (1994).Towards a new economics of science, Research Policy, 23, 487-507.

David, P.A., and Foray, D. (2003). Economics fundamental of the Knowledge Society. Policy Futures in Education 1(1), 20–49.

Dorogovtsev, S.N., and Mendes, S.N. (2003). Evolution of Networks: From biological nets to the internet and www. Oxford: Oxford University Press.

Eckmann J.P. and Moses E (2006). Curvature of co-links uncovers hidden thematic layers in the World Wide Web. Proceedings of the National Academic of Sciences of the USA, 99, 5825-5825.

Egghe, L., and Rousseau, R. (1990). Introduction to informetrics: Quantitative methods in library, documentation and information science. Amsterdam: Elsevier Science Publishers.

Freeman, L.C. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40, 35-41.

Garfield, E. (1972) Citation analysis as a tool in journal evolution. Science, 178, 471-479.

Garfield, E., Pudovkin, A.I., and Istomin, V.I. (2003). Mapping the output of topical searches in the Web of Knowledge and the case of Watson-Crick. Information Technology and Libraries, 22(4), 183–187.

Geller, N.L. (1978). On the Citation Influence Methodology of Pinski and Narin. Information Processing and Management, 14, 93-95.

Gibbons, M., Limoges C., Nowotny, A., Schwartzman, S., Scott P., and Trow M. (1994). The New Production of Knowledge: The Dynamics of Science and research in Contemporary Societies, Sage, London.

Girvan, M. and Newman, M.E.J. (2002).Community structure in social and biological networks. Proceedings of the National Academic of Sciences of the USA 99, 7821-7826.

Harary, F. (1996). Graph Theory. Cambridge, MA: Perseus.

Hauser, J., and Katz, G. (1998): Metrics: You are what you measure, European Management Journal, 16(5), 517-528.

Hummon, N. and Carley, K. (1993). Social networks as normal science. Social Networks, 15, 71–106.

Hummon, N. and Doreian, P. (1989). Connectivity in a citation network: the development of DNA theory. Social Networks, 11, 39–63.

Hummon, N. and Doreian P. (1990). Computational methods for social network analysis. Social Networks, 12, 273–88.

Katz, J.S. and Martijn, B.R. (1997) What is research collaboration?. Research Policy, 26 (1), 1-18.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. Journal of the Association for Computing Machinery, 46 (5), 604-632.

Kuhn, T. (1962). The Structure of Scientific Revolutions. Chicago: Chicago University Press.

Lambiotte, R., P. Panzarasa (2009). Communities, Knowledge Creation and Information Diffusion. In: Katy Börner, Andrea Scharnhorst, (Eds.), Journal of Informetrics, Special Issue on the Science of Science: Conceptualizations and Models of Science, 3(3), 180–190.

Lawani, S.M. (1986) 'Some bibliometric correlates of quality in scientific research', Scientometrics, 9, 13–25.

Lusseau D. and Newman, M.E.J. (2004) Identifying the role that animals play in their social networks. Proceedings of the Royal Society of London B 271 S6: S477-S481.

Martin, B.R., and Irvine, J. (1983). Assessing basic research. Some partial indicators of scientific progress in radio astronomy. Research Policy, 12, 61-90.

Martin-Sempere, M. J., Rey-Rocha, J., and Garzon-Garcia, B. (2002). The effect of team consolidation on research collaboration and performance of scientists. Case study of Spanish University researchers in Geology. Scientometrics, 55(3), 377–394.

Melin, G., and Persson O. (1996). Studying research collaboration using co-authorships, Scientometrics , 36, 363-377.

Milgram, S. (1967). The small world problem. Psychology Today 2, 60-67.

Mina, A., Ramlogan, R., Tampubolon, G. and Metcalfe, J.S. (2007). Mapping Evolutionary trajectories: Applications to the growth and transformation of medical knowledge. Research Policy, 36, 789-806.

Moed, H.F., De Bruin, R.E., and Van Leeuwen, T.N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. Scientometrics, 33, 381-425.

Narin, F., and Whitlow, E.S. (1990). Measurement of scientific co-operation and co-authorship in CEC-related areas of science, Report EUR 12900, Office for Official Publications of the European Communities, Luxembourg.

Nederhof, A.J., and Visser, M.S. (2004). Quantitative deconstruction of citation impact indicators: Waxing field impact but waning journal impact. Journal of Documentation, 60 (6), 658-672.

Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. Scientometrics, 66, 81-100.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. PNAS, 98 (2), 404-409.

Newman, M. E. J. (2002). The structure and function of networks, Computer Physics Communications, 147, 40–45.

Newman, M.E.J. (2004). Detecting community structure in networks. The European Physical Journal B, 38(2), 321–330.

Newman, M.E.J. and Girvan, M. (2004) Finding and evaluating community structure in networks. Physical Review E 69, article no. 026113.

Newman, M.E.J. (2008). The physics of networks. Physics Today, November 2008, 33-38.

Newman, M.E.J. (2010). Networks: An Introduction. Oxford University Press.

Newman, M.E.J., Barabási, A., and Watts D. (2006). The Structure and Dynamics of Networks. Princeton University Press.

Noyons, E.C.M. (1999). Bibliometric Mapping as a science policy and research management tool. Thesis Leiden University. Leiden: DSWO Press.

Noyons, E.C.M., Luwel, M. and Moed, H.F. (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes. Journal of the American Society for Information Science 50, 115-131.

OECD, The Knowledge-Based Economy, Paris, 1996.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web (Tech. Rep.) Stanford Digital Library Technologies Project.

Persson, O., and Beckman M. (1995). Locating the network of interacting authors in scientific specialties, Scientometrics, 33, 351-366.

Peters, H.P.F. and van Raan, A.F.J. (1994). On determinants of Citation Scores-A-Case-Study in Chemical Engineering. Journal of the American Society for Information Science, 45(1), 39-19.

Pillai, S.U., Suel, T., and Cha, S.H. (2005) The Perron-Frobenius theorem: some of its applications. IEEE Signal Processing Magazine, 22(2), 62–75.

Pinski, G. and Narin, F. (1976). Citation Influence For Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics. Information Processing and Management, 12, 297-312.

Pool, I. de S., and Kochen, M. (1978). Contacts and Influence. Social Networks 1, 1-48.

Ramlogan, R., Mina, A. Tampubolon, G., and Metcalfe, J.S. (2007). Networks of Knowledge: The Distributed Nature of Medical Innovation. Scientometrics, 70 (2), 459-489.

Rinia, E.J., Van Leeuwen, T.N., Van Vuren, H.G., and Van Raan, A.F.J. (2001). Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research. Research Policy, 30 (3), 357-361.

Rosvall, M. (2006). Information Horizons in a Complex World, Ph.D. Thesis; Umeå University, Umeå.

Seglen, Per O., and Aksnes Dag W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research, Scientometrics, 49 (1), 125-143.

Seiman, S. (1983). Network structure and minimum degree, Social Networks, 5, 269-287.

Scott, J. (2000). Social Network Analysis: A Handbook. Newbury Park, CA: Sage Publications.

Schubert, A., Glänzel, W., and Braun, T. (1989). Scientometric data files. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981-1985. Scientometrics, 16, 3-478.

Schmoch, U., Schubert T., Jansen D., Heidler R., and von Görtz, R. (2010) How to Use Indicators to Measure Scientific Performance? A Balanced Approach, Research Evaluation, 19 (1), 2-18.

Tijssen, R., Hollanders, H., Van Leeuwen, T., and Nederhof, A.J. (2006). Science and Technology indicators 2005 Summary. Netherlands Observatory of Science and Technology. Den Haag: Deltahage.

Tijssen, R.J.W., Van Leeuwen, T.N., and Van Raan, A.F.J. (2002). Mapping the scientific performance of German Medical research. Stuttgart, New York: Schattauer.

Van Eck, N.J. and L. Waltman (2007). Bibliometric mapping of the computational intelligence field. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15(5), 625-645.

Van Eck, N.J., and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538.

Van Leeuwen, T.N, (2004). Second generation bibliometric indicators. The improvement of existing and development of new bibliometric indicators for research and journal performance assessment procedures. Thesis Leiden University. Leiden: DSWO Press.

Van Leeuwen, T.N., Visser, M.S., Moed, H.F., Nederhof, A.J., and Van Raan, A.F.J. (2003). The holy grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. Scientometrics, 57, 2, 257-280.

Van Raan, A.F.J. (1997). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientometrics, 36, 396-420.

Van Raan, A. F. J. (1998). The influence of international collaboration of the impact of the research results. Some simple mathematical considerations concerning the role of self-citations. Scientometrics, 42(3), 423–428.

Van Raan, A.F.J. (2004). Measuring Science. Capita Selecta of Current Main Issues.In: H.F. Moed, W, Glänzel and U. Schmoch, editors: Handbook of Quantitative Science and Technology Research, Dordrecht: Kluwer Academic Publishers, 2004, p.19-50.

Van Raan, A.F.J. (2008). Scaling Rules in the Science System: Influence of Field-Specific Citation Characteristics on the Impact of Research Groups. Journal of the American Society for Information Science and Technology, 59 (4), 565-576.

Van Raan, A.F.J., and Noyons, E.C.M. (2002) Discovery of patterns of scientific and technological development and knowledge transfer. In: Adamczak, W., Nase, A. (eds.): Gaining Insight from Research Information. Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29-31. Kassel: University Press, p. 105-112.

Wasserman, S. and Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Watts, D.J. (1999). Small Worlds: The Dynamics of Networks between Order and Randomness, Princeton University Press, Princeton.

Watts, D.J., and Strogatz, S. (1998). Collective dynamics of small-world networks. Nature, 393, 440-442.

Wutchy, S., Jones, B.F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316, 1036-1039.

2 How to identify research groups using publication analysis: an example in the field of nanotechnology

Clara Calero , Renald Buter , Cecilia Cabello Valdés, Ed Noyons
(published in Scientometrics 66(2): 365-376, 2006)

2.1 Introduction

On a conceptual level, much has been made of the observed switch from "Mode 1" to "Mode 2" models or types of research and knowledge generation put forward by Michael Gibbons and co-workers. The model shift has been related with the trend towards multi- and inter-disciplinary research and the long term decline of single discipline research, but also in the increased wish and need for collaboration in researchers. (PREST 2000). This approach recognizes that research is a collective effort, combining diverse actors, competences and capabilities. It puts the emphasis upon the collective setting, intermediary between individual researchers and research institutions (Laredo 2003). Although the typical research group still has a core team consisting of tenured staff and students (graduate, doctoral and postdoctoral), there is usually a more peripheral level of visiting scientists and cooperating domestic and foreign colleagues. And actually are these broad cooperative elements the actual research-performing units, which may reflect the realities of the scientific process more accurately that do core teams. (Seglen and Aksness 2000). In such a framework, policies/strategies cannot rely only on a content dimensions i.e.thematic priorities, they have also to care about organizational aspects. Questions such as: Do we have the right research groups? Are they inter-connected enough? What about their connections with their environment? Are more and more pressing (Laredo 2003)?

It has been stated often that bibliometric analyses could play an important role in measure these tendencies and along this road a number of studies have been carrying out lately. Recently many improvements have been made in getting an overview of multi-and inter-disciplinary fields through bibliometric maps (Noyons 1999, Noyons et al. 2002, and Noyons et al. 2003). The co-authorship data were used in many studies to measure collaboration (eg. Persson & Beckmann 1995, Melin & Persson 1996, Bordons & Gomez 2000, Seglen & Aksness 2000) and starting around 2000 several researchers began the construction of large-scale networks using co-authorship data representing research in mathematics (Barabási et al 2002); biology, physics and computer science (Newman 2001); and neuroscience (Barabási et al 2002). However, most of these studies are fragmented, focusing on one or a few characteristics of the process at a time. Only a few attempts have been made to relate cognitive structures and collaboration (eg. Mutsche & Quan Haase 2001). Here, we used bibliometric mapping techniques and network analysis to identify and classify research groups. This approach intends to cover the two key trends of the knowledge process: multi-and inter-disciplinary scientific fields and broad cooperative units of research.

The aim of this paper is to present a new approach to identify research groups analyzing the articles published in scientific journals in a particular science field.

2.2 Data and Methods

The data for this study were taken from a project financed by the Spanish Foundation for Science and Technology (FECYT). One of the objectives of the project was to map and identify Spanish research groups in the field of nanotechnology.

Delineation (publication data collection)

The data collection (or delineation) procedure was carefully designed in close collaboration with the field experts. In a first step, core publications for the field were collected. The database Current Contents from the Institute for Scientific Information (ISI) was used as a source of primary data. This primary collection was based on the delineation adopted in the EC mapping of excellence project (Noyons et al 2003). We took the final discussion from that project into consideration and compiled a primary search strategy for the FECYT project. From this core set of publications we extracted candidate search terms to expand the set of publications and asked experts to indicate the relevant candidates (or suggest alternatives). The FECYT and the expert groups involved in this project considered that for the delineation of this field special emphasis should give to materials, because the importance they have in Spain. In a second round the suggested terms were used and a new data were collected. The results in this study were based on this second search strategy.

The core publications for the field were collected by the following search terms:

- nano* NOT (nanomet* OR nano2 OR nano3 OR nano4 OR nano5 OR nanosecon* OR nano secon*) OR
- nanomet* scale* OR nanometerscale* OR nanometer length OR nano meter length
- nanoa* OR nanob* OR nanoc* OR nanod* OR nanoe* OR nanof* OR nanog* OR nanoh* OR nanoi OR nanoj* OR nanok* OR nanol* OR nanon* OR nanoo* OR nanop* OR nanoq* OR nanor* OR nanot* OR nanou* OR nanov* OR nanow* OR nanox* OR nanoy* OR nano z*
- atom* force microscop*

- tunnel* microscop*
- scanning probe microscopy
- scanning force microscop*
- semiconductor quantum dot
- silicon quantum dot
- quantum dot array
- coulomb blockade
- Single molecule
- molecular motor
- molecular beacon
- biosensor
- self-organized growth
- electron beam lithography
- monolayers growth
- optoelectronic* device*
- Quantum Computing
- quantum devices
- quantum Discs
- quantum optoelectronics
- quantum Wells
- quantum wires
- Scanning probes techniques
- Transmission electron microscopy
- resonant cavity
- resonant cavities
- self assembling
- self ordering
- spintronics
- submicron devices
- vertical cavity surface emitting Laser*
- cantilevers
- quantum dots
- Molecular Beam Epitaxy

The final set from the period January 1996- January 2003 contained a total of 91,372 articles retrieved from the above mentioned database Current Contents.

Bibliometric Mapping

The bibliometric map is a two-dimensional representation of the core publications collection, designating the field of nanotechnology. From these publications, we extracted noun phrases from titles and abstracts to be used for a co-occurrence analysis. This co-occurrence analysis clusters selected noun phrases (keywords). These keywords were identified from the endless list of noun phrases, on the basis of bibliometric distribution, syntactic features and (semantic) content.

The clusters of keywords designated sub-domains in the field. By these keywords, publications were assigned to sub-domains. Thus, sub-domains were in fact, subsets of publications from the entire collection, the field. As publications might be assigned to more than one sub-domain, we generated a co-occurrence matrix of sub-domains. The cells in this N*N matrix (in which N designates the number of sub-domains), contained the number of publications overlapping in two of the N sub-domains. This matrix was the input for Multidimensional scaling (MDS). This technique put the N elements in a two-dimensional space in such a way that sub-domains with a similar orientation in relation to all other the sub-domains, were in each other vicinity, whereas sub-domains with a different orientation were distant from each other. This two dimensional representation was the bibliometric map of the field.

Figure 1 shows the bibliometric map of the field of nanotechnology for the present study.

Identification of a research group

The analysis was based on units formed by combinations of author name and main organization. In this case, because of the scope of the project, all the organizations selected are from Spain. The research groups were identified and defined on the basis of similar research activity profiles and co-authorship.

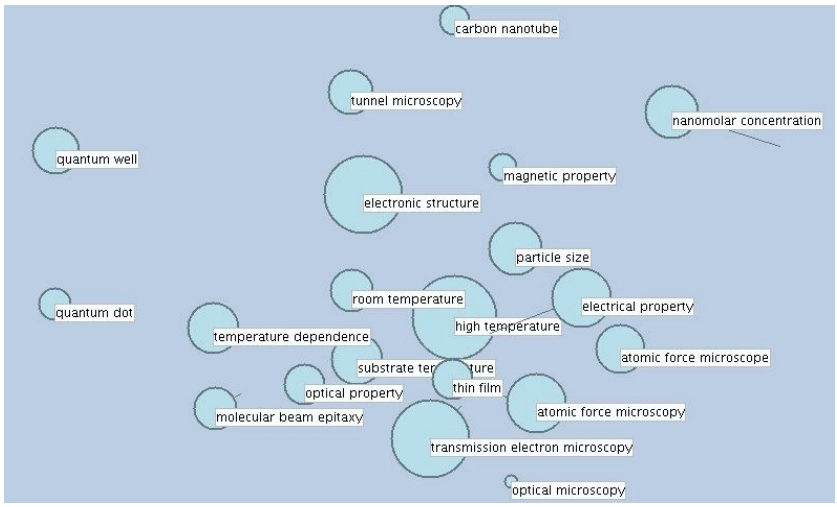

Figure 1. Bibliometric map of Nanotechnology

The sub-domains (clusters of topics) are positioned, depending on the cognitive orientation. The more two sub-domains are related the closer they are. Each sub-domain is characterized by the most prominent keyword within. The size of the surface indicates the number of publications represented.

Author/Organization Combination (AOC)

We assumed that we could define a group bibliometrically by a collection of publications. This collection was identified by the oeuvres of one or a set of authors. In order to do that, we had to deal with the publication author names. We encountered two problems in publications data related with the author's field: two persons with the same author name (homonymous names) or two or more author names referring to the same person (synonymous names).

To solve the problem with homonymous names, we used a combination of author names and main organization (university, company…). Each publication had in most cases at least one author and at least one address (from the address field it was considered just the organization); in this case the only thing we knew was that the first author is attached to the first organization. But as the first author may also be at the second or third organization and the second author may be attached to any of the organizations… We assigned in a publication all author names to all organizations. So for a publication with 3 authors and 2 organizations[1],

[1] Actually in the publication itself the author name is attached to the organization/s that belongs to. This doesn't happen on the information contained on the electronic databases.

we defined in fact 6 (3*2) authors. Or more correctly, we define 3 authors associated with two organizations.

Example of AOCs

Publication X		AOCs in Publication X
Authors		• A 1, Org A
1 A 1		• A 2, Org A
2 A 2	=>	• A 3, Org A
3 A 3	*3x2*	• A 1, Org B
Addresses		• A 2, Org B
• Org A		• A 3, Org B
• Org B		

This solution, however, increased the problem of the second type (synonymy). As each author was associated with all the organizations in a publication, more names were referring to the same person. The analysis of the relations between the AOCs dealt with this problem.

Because the purpose of this study was to identify Spanish research groups, we selected only the organizations coming from Spain. Besides, only AOCs with more than six publications were considered.

Activity Similarity Relations

Using the bibliometric mapping and clustering analysis of the field of nanotechnology we created the research profile of each AOC. In our database this profile was compiled by the number of publications the AOC had in each cluster.

The next step was to compare the AOCs on the basics of this activity profile. In order to do so we used a similarity measure, the cosine coefficient (Noyons 1999). Each pair of AOCs got a value between 0 and 1 indicating their similarity. Because the objective of our study was to identify research groups based on their research activity similarity, we considered only the relations with a cosine coefficient higher than 0.9[2].

[2] The threshold of 0.9 was arbitrary, but a small test using other thresholds did not yield significantly different results.

Co-publication Relations

We constructed the co-author/organization matrix composed by co-occurrences of the AOCs co-publishing the same article.

Network Analysis

A network analysis was applied to represent the two different relations explained above and to identify community structures. In the network theory the graphs are composed of nodes (or actors or points or vertices) connected by edges (or relations or ties).

For the identification of subgroups of authors within the network, we used the k-core approach. A k-core is a subgraph in which each node (AOC) is connected to at least a minimum fixed number (K) of the other nodes in the subgraph. The k-core approach is less strict (compared with others like cliques, n-cliques, n-clans…), allowing actors to join the group if they are connected to k members, regardless of how many other members they may not be connected to. (Wasserman and Faust 1994).

So the Activity Similarity Graph represented the relations between AOCs connected by similar research activity profile, while the Co-author Graphs represented the relations between AOCs by the absolute number of co-publications.

The Activity Similarity Graph was used as a base for the analysis. We extracted the subgraphs from this network. Each subgraphs extracted was analyzed also using the co-authorship.

2.3 Results

The results expose in this section are two examples of the subgraphs extracted from the activity similarity graph and illustrate the application and potential of this new methodology. These cases are representative of the two types of outcomes expected from this method: the identification of a research group and the detection of potential partners.

In the Activity Similarity Graph each node represents an AOC and the relations depict a similar research profile (Figure 2). This graph is the starting point to identify subgroups of AOCs. As mentioned in the previous section we used the K-core approach to divide this network into subgraphs.

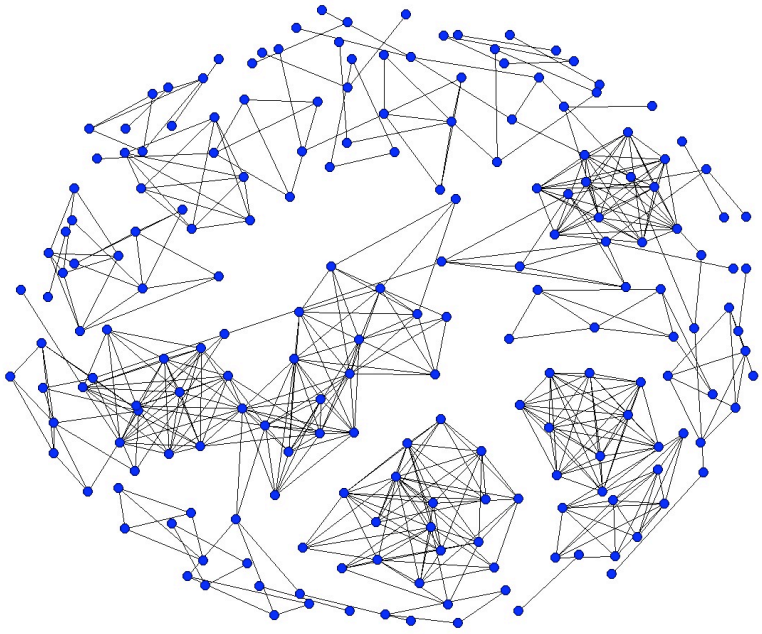

Figure 2. Activity Similarity Graph of Nanotechnology

Each node represents an author/organization combination (AOC). A connecting line indicates a similar research profile.

Identification of a research group

The Figure 3 shows a subgraph of ten AOCs; each of them is related with, at least, eight of the others. We have identified already a group of AOCs with a very similar research profile. Consequently this is a potential research group, but are these AOCs working together? The Figure 4 illustrates the connections between these ten AOCs based on their co-authorship. As we can see these AOCs are actually working together. So they are a research group. The last step it will be to assign the AOCs to the authors and organizations that they are related with.

If we take a closer look to the individual AOCs, we can see that for each author name there are two organizations related: Cádiz University and Polytechnic University of Madrid (UPM). But we are not concern about an author or an organization we are looking for a group. With the information contained in figure 3 and 4, we have identified a research group composed of six researchers: Izpura I, Gutiérrez M , Aragón G,

González D, García R, and Sánchez JJ; coming from two universities: Cádiz University and Polytechnic University of Madrid.

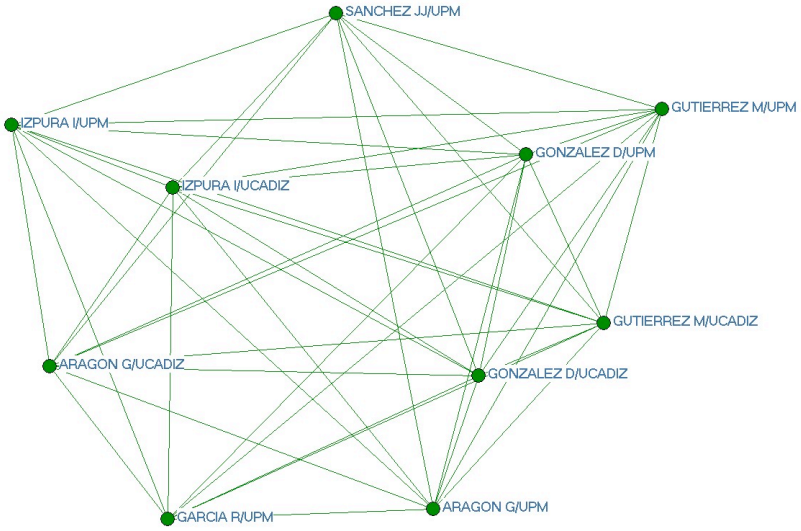

Figure 3. 8-core subgraph based on activity similarity relations.
Each pair of AOCS connected represents a similar research profile.

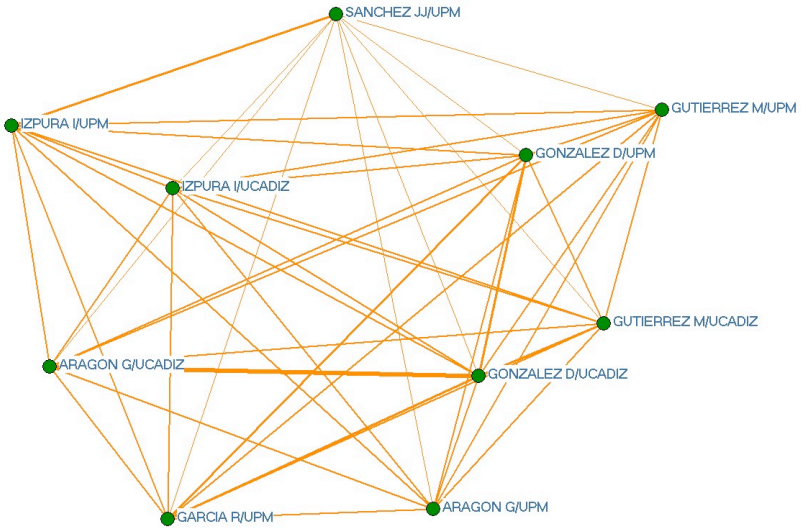

Figure 4. Subgraph based on co-authorship relations.
Each pair of AOCS connected shows a co-published activity.

Identification of potential research partners

The second example illustrates other sub-graph extracted from the Activity Similarity Graph (Figure 5). In this case each AOC is connected with at least three of the others. As the previous example, the relations mean a similar research profile. Again, we can take a look in Figure 6 to the co-authorship relations. Here, we get two co-publications groups. At the top part of the Figure 6, there is a co-author group coming from Alicante University formed by four researchers: Herrero E, Rodes A, Feliú JM, and Gomez R. The second co-author group belongs to the Autonomous University of Madrid (UAM) and to the CSIC-ICMA (one of the institutes at the Spanish Council for Scientific Research) formed by three researchers: Gómez Herrero J., Ordejón P. and Baró AM. Finally, there are two AOCs not co-publishing in this set: Agrait N/UAM and Flores F/UAM.

The activity similarity subgraph has identified a group of AOCs with similar research profile. While the coauthor data depicts that this set is divided into two co-authors groups and two isolated AOCs. The information provided by the activity similarity subgraph identified potential partners for the AOCs that not co-publish.

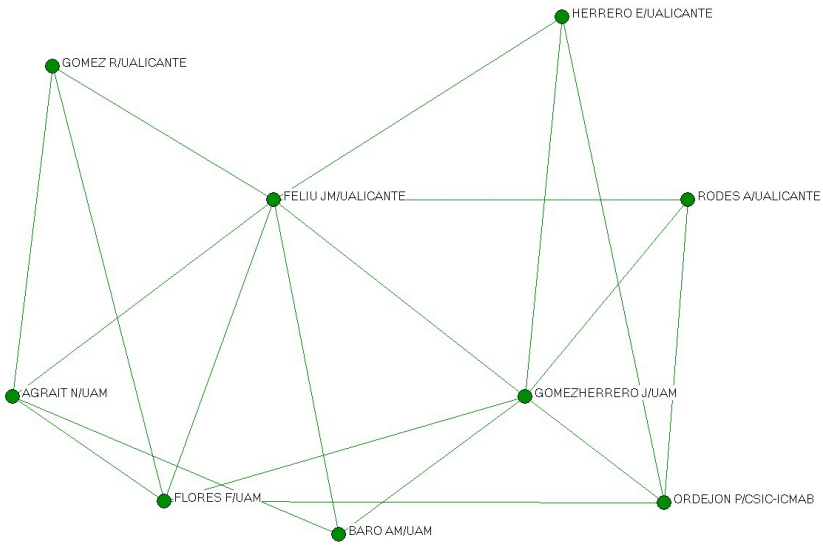

Figure 5. 3-core subgraph based on activity similarity relations.
Each pair of AOCS connected represents a similar research profile.

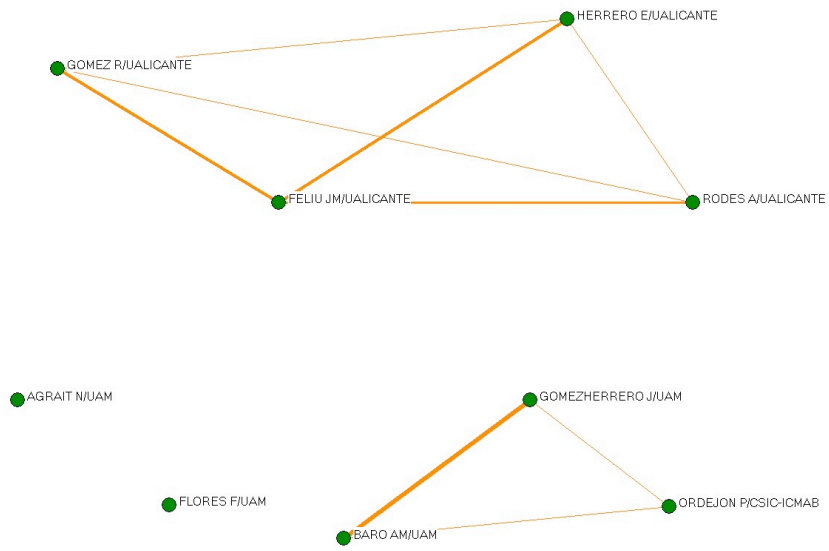

Figure 6. Subgraph based on co-authorship relations.
Each pair of AOCS connected shows a co-published activity.

2.4 Conclusions and Discussion

In the last years we have observed considerable advances in measuring the knowledge production and utilization. The idea that the scientific research is moving from a personal, disciplinary-based, and place-bound ideal towards a collective, problem-oriented and multi-organizational activity is well-accepted nowadays.

The method presented here should be considered only as the starting point toward a complete methodology for identifying research groups and potential research partners in scientific fields. A first and important result of the study regards with the matter that we have identified functional rather than physical groups. Following Seglen and Aksness (2000) definition of a research group: "…a research group assignment based on co-authorship defines functional rather than physical groups, and might include, e.g. authors with whom a group member has collaborated in connection with a short-term scientific visit. Our group concept is thus somewhat wider and loser than the standard conception of a physically localized research team". The groups are defined over a six year time period meaning that the group members have not necessarily worked together. In addition, the identification of the members through the AOCs

allows the same person to belong to more than one group. This is the case, for instance, of a researcher that moves from organization and changes his line of research.

A second significant outcome of the study concerns the idea of being able to identify potential research partners. Using the activity similarity relations combined with the co-author relations it is possible to detect groups working on the same areas but not co-publishing.

A third important result of our approach is that we should be able to deal with the homonymous and synonymous names. The combination of the author and the address field in a publication allow us to solve the problem of the homonymous names, while the network analysis provides a possibility to deal with the latter. The combined data enables us to assign more accurately author names to authors**.**

Nevertheless, we are aware that there are a number of potential improvements that can be made to the method presented here, including the following suggestions to be implemented in further research:

- Validate the results with the opinions of the experts in the field.
- Analyze in more detail the profile and position of some authors in the activity similarity network to identify authors that are 'bridges' between groups with different profiles.
- Add to the analysis the impact factor for each AOC to identify success teams.
- Use other techniques related with community structures in networks and compare with the results from the K-core approach.
- Enlarge the scope of the analysis to international collaborations.
- Another issue to consider is the time evolution of the identified groups.

In summary, the method and results presented here should be considered a starting point for developing a methodology to identify systematic research groups. It is important to note that such method is open: more details are going to be incorporated which are going to improve the results.

References

Barabási, A.-L., Jeong, H., Ravasz, E., Néda, Z., Schuberts, A., and Vicsek T. (2002), Evolution of the social network of scientific collaborations, Physica, A 311: 590–614.

Bordons, M., and Gómez I. (2000), Collaboration networks. in science. In: H. B. Atkins, B. Cronin (Eds.), The Web of Knowledge: A Festschrift in Honor of Eugene Garfield, Information Today, Medford, NJ, pp. 197-213.

Gibbons, M., Limoges C., Nowotny, A., Schwartzman, S., Scott P., and Trow M. (1994), The New Production of Knowledge: The Dynamics of Science and research in Contemporary Societies, Sage, London.

Laredo, P. (2003), University research activities On-going transformations and new challenges, Higher Education Management and Policy, 15 (1): 138 – 163.

Melin, G., and Persson O. (1996), Studying research collaboration using co-authorships, Scientometrics , 36: 363-377.

Mustchke, P., and Haase, A.Q. (2001), Collaboration and cognitive structures in social science research fields. Towards socio-cognitive analysis in information systems, Scientometrics, 52 (3): 487-502.

Newman, M.E.J. (2001), The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98, 404-409.

Noyons, E.C.M. (1999), Bibliometric mapping as a science policy and research management tool, Thesis Leiden University. Leiden: DSWO Press (ISBN 90-6695-152-4).

Noyons, E.C.M., Buter R.K., and van Raan A.F.J. (2000), Mapping the field of Neuroscience. Electronic version with interactive facilities available via www.cwts.leidenuniv.nl.

Noyons, E.C.M.,. Buter, R.K., van Raan, A.F.J., Schmoch, U., Heinze, T., Hinze S., and Rangnow R. Mapping Excellence in Science and Technology across Europe: Nanoscience and Nanotechnology. Report of project EC-PPN CT 2002-0001 to the European Commission, Leiden, October 2003, 113 pp.

Persson, O., and Beckman M. (1995), Locating the network of interacting authors in scientific specialties, Scientometrics, 33: 351-366.

PREST (2000) Impact of the Research Assessment Exercise and the Future of Quality Assurance in the Light of Changes in the Research Landscape, prepared for Higher Education Funding Council for England (HEFCE), April 2000, available from www.hefce.ac.uk.

Seglen, Per O., and Aksnes Dag W. (2000), Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research, Scientometrics 49 (1): 125-143.

Wasserman, S., and Faust K. (1994), Social Network Analysis, Cambridge University Press, Cambridge.

3 Reorganizing research with the help of bibliometric collaboration networks. Case study in a University Hospital.

Clara Calero Medina and Thed N van Leeuwen
(submitted)

3.1 Introduction

Interdisciplinarity and collaboration

Bibliometric analyses nowadays are more and more focussed on subjects that are relevant for contemporary topics in science policy and management. One of the topics that have gained relevance in the bibliometric field in the passed years is the evaluation and study of interdisciplinary research. The growing importance of research at the boundaries of existing traditional disciplines during the last years demands for further validation of existing methods, and the development of new methods to study and evaluate interdisciplinary research fields (Rinia, 2000).

There are many different scientometric approaches to measuring interdisciplinary, each relying both on a system of disciplinary categories and a concept of interdisciplinarity (Schummer, 2004). Many approaches take papers (or patents) as the subject of study and measure interdisciplinarity in terms of the co-occurrences of what can be considered discipline specific items, such as main concepts (Borner, K., Chen C.M., & Boyack K.W., 2003, and Van Eck & Waltman, 2007), classification headings (eg. Steele & Stier, 2000, Morillo, Bordons & Gómez, 2001, and Rinia et al., 2001), authors' affiliations (Qin, Lancaster, & Allen, 1997, and Steele & Stier, 2000), or citations (eg. Bourke & Butler, 1998, Steele & Stier, 2000 , Rinia et al., 2002, and van Raan & Leeuwen, 2002).

In this study, we are interested in interdisciplinarity as a combined cognitive and social phenomenon (Schummer, 2004), which is particular important in highly interdisciplinary environments as a university hospital. Levels and types of interdisciplinary collaboration vary in different disciplines, but the general trend is toward increasing interdisciplinarity, which is particularly pronounced in biology and medical sciences (Qin, Lancaster, & Allen, 1997).

In interdisciplinary research environments, groups with in principle different research interests and backgrounds, collaborate pooling their knowledge toward a common goal (Qin, Lancaster, & Allen, 1997). This collaboration across disciplinary boundaries is a more complex form of information transfer (Pierce, 1999). Gibbons et al. (1994) have described it as a "transdisciplinary" form of collaboration with members of different disciplines working together in practical applications. The mutual, direct engagement among previously uncorrelated research topics has advantages not only for the researchers, that are able to draw from a

wider, diverse intellectual environment, but also for the nature of research performed, that is circulated, validated and enriched by contact with new research and social circles (Pierce, 1999). This is why large interdisciplinary research centers are interesting environments to study collaboration (Rodriguez & Pepe, 2008).

One way to analyze and study the ways researchers exchange and share information is through the construction of co-authorships networks (Newman, 2001; Barabási et al., 2002; Newman, 2003). Network researchers have been working during the last decade to reveal the highly clustered nature of scientific production, showing that co-authorships and citation networks are made of several dense groups of nodes, called communities (Lambiotte & Panzara, 2009). Allowing the co-publication network itself to identify communities and groups, what we call *functional research groups*[1] (Seglen & Aksnes, 2001; Calero et al., 2006) inside interdisciplinary research centers, may lead to a better understanding of how these complex organizations work and therefore can help research managers to reorganize the organization in a more efficient and practical way.

Main Characteristics of the Leiden University Medical Center (LUMC) in The Netherlands

The LUMC has a long tradition of pioneering medical and bio-medical research and is among the international top in this field. With its research, the LUMC wants to contribute to the prevention and solution of health problems. At the heart of their research strategy is translational research. Translational research in medicine means the effective translation of the new knowledge, mechanisms, and techniques generated by advances in basic medical research into new approaches for prevention, diagnosis, and treatment of diseases (Fontanerosa & DeAngelis, 2002; Woolf, 2008). This is a prime example of how an interdisciplinary approach can promote the use of clinical and laboratory findings in applications that are beneficial to society and citizens. The LUMC has structured its research in *Departmental Programmes* and *Research Themes.*

Almost all LUMC departments do conduct research in addition to their other tasks (teaching, patient care). This research is mainly structured in the form of *Departmental Programmes* (see **Table A** in the Appendix). The *Departmental Programmes* follow the lines of traditional research fields and there are about 150 of them. *Departmental Programmes* are

[1] The functional research groups are broad cooperative units of research identified through co-authorship activity, not necessarily embedded in the traditional physical groups of the organization..

representative for the scientific activities of the LUMC. But the *Departmental Programmes* do not stand-alone; there is a high degree of cohesion between the various *programmes*.

To intensify and to profile scientific collaboration in the LUMC, in 2006 a more structured approach was introduced in the form of research themes. A research theme is an intra-*Departmental* or cross-divisional collaborative alliance of researchers who jointly study a single topic, each of them on the basis of their own disciplines. *Research themes* focus on a particular illness or clinical picture. A theme is intended to promote synergy. The theme leader is accountable to a division governing board, which is ultimately responsible for the theme. The LUMC *Research Themes* are:

- Aging.
- Genetic Epidemiology and Bioinformatics.
- Immunotherapy of cancer.
- Infectious diseases and immunology.
- Neurosciences.
- Oncogenetics.
- Regenerative Medicine.
- Vascular Medicine.

3.2 Objectives of this study

In the Netherlands, research performance assessments are often extended with bibliometric analyses. In the medical sciences, the Royal Academy of Sciences (KNAW) has long been the initiator for research assessments in the field. In 2003, a new evaluation protocol, the Standard Evaluation Protocol (SEP) was implanted in the Dutch science system. In this new protocol, the responsibility and initiative for research evaluation is transferred to the individual Boards of Dutch universities.

Under this new protocol, the Board of the Leiden University Medical Center (LUMC) has initiated an annual bibliometric monitoring of the research performance of the research within the LUMC. The Board is aware of the importance for a highly interdisciplinary environment as LUMC of a proper research management to facilitate the research activities. The network of interactions (co-authorships) is considerably more complex than shown through the formal organization (*Departmental Programmes* and *research themes)*. LUMC expects that the bibliometric monitoring will help them, among other issues, to change

the configuration of the *Departmental Programmes* orienting them towards *Research Themes*, in order to create as strong as possible research clusters.

In terms of policy issues, the question raised by the LUMC Board was to identify the "functional research groups" inside LUMC that could help them to re-organize their research lines (*Departmental Programmes* and *Research Themes*).

3.3 Methodology

Data

Publications from 2002 until 2006 were extracted from an in-house LUMC output registration system and matched with Web of Science (WoS). The resulting dataset, containing publications labeled with LUMC *Departmental Programmes* names and *Research Themes* names forms the basis for the first and second analysis respectively.

Departmental Programmes Analysis

For the analysis of the *Departmental Programmes* first the LUMC's publication set is classified in research fields. For the most important fields in terms of number of publications, we collect which *Departmental Programmes* are publishing in each of the fields selected and their co-publication activity. We present a detailed explanation below.

- *Inverse Research Profile*

The *breakdown* of the LUMC output (publications) into research fields (our definition of research fields is based on the classification of scientific journals into Journal Subjects Categories developed by Thomson Reuters) is what we call *research profile*. The break down as such gives an impression of all fields involved in the research scope or 'profile' of the LUMC. This can be seen already as an indicator of interdisciplinarity (Van Raan & Van Leeuwen, 2002). Additionally, we determine the *impact* of the articles in these fields, so it becomes immediately visible in what fields within the 'research profile' the LUMC has a high (or low) performance (Moed, De Bruin & Van Leeuwen, 1995). In terms of the policy use of the profiles, the impact is compared to the mean field citation score (CPP/FCSm[2]).

[2] **CPP** is the average number of citations per publication (excluding self-citations)
FCSm is the reference value. The average citation rate of all articles in the subfields

Having as starting point the overall 'research profile' of the LUMC, in which the fields are presented in such a way that the largest research fields are on top of the profile, we get a new insight in the (WoS)information available. Per research field, we have collected the 'research profile' information of the *Departmental Programmes*, and displayed the information in an inverse way: the *Departmental Programmes* with the largest output are on top of the profile, indicating their relative strong contribution to the field, again in combination with the impact received in that field. This is what we call '*inverse research Departmental Programmes profiles'*. Only research fields with more than 3% of all publications by the LUMC were considered. As a result, a total of 10 fields were considered and inverse research *Departmental Programmes* profiles were calculated for them.

- *Field collaboration Departmental Programmes: network analysis*

The various *Departmental Programmes* of the LUMC work together in many ways. Using each of the 'inverse *Departmental Programmes* profiles' per field, an analysis was conducted on the scientific cooperation relationships among the *Departmental Programmes*. Each of the 10 collaboration networks is then a set of *Departmental Programmes* (in network theory called nodes) linked via their co-publication activity (edges). The network is undirected which means that the edges have no direction. The edges are valued representing the strength of the co-publications activity between two *Departmental Programmes*. The collaboration network analysis is focused on the structures per field within which the *Departmental Programmes* are embedded. It is intended to create a better insight in the visibility and relatedness of the *Departmental Programmes*.

In network analysis there is a number of techniques to detect the cohesion and cohesive subgroups inside the network (Scott, 2000). Intuitively, cohesion means that a network contains many ties. More ties between vertices yield a tighter structure, which is, presumably, more cohesive (De Nooy, Mrvar & Batagelj, 2005). These techniques are based on the way in which vertices are interconnected, in our case in the way in which

(Thomson Reuters Journal Subject Categories) in which the research unit analyzed is active (excluding self-citations) Also indicated as the world citation average in those subfields or 'world subfield average'. Then the **CPP/FCSm** is the impact of a research unit's articles, compared to the world citation average in the subfields in which the research unit is active.

the research *Programmes* are interconnected in their collaboration activity. We expect that the LUMC *Departmental Programmes* with a common research activity will interact frequently, at least more than with other *Programmes*.

First we are interested in general characteristics and properties of the collaboration networks, such as:

❖ *Size*

The size of the network is measured in term of the number of research *Programmes* involved.

❖ *Density and Clustering*

The density describes the general level of linkages among the research *Programmes* in the collaboration network. The more the research *Programmes* are connected to one another, the more dense the network will. A 'complete network' is one in which all the research *Programmes* are connected to one another. The *Departmental Programmes* collaboration network is a valued graph. The density measure with valued graphs is complicate to calculate and interpret, especially because it is highly sensitive to the assumptions that we have made about the data (Scott, 2000). This is why we are calculating the density disregarding the values of the lines. We are aware that this involves a considerable loss of information, but at the same time it gives a first insight of the cohesion of the network. We will complement this measure with the clustering coefficient in order to get as much as information as possible.

The maximum number of edges for a network is determined by

$$E_{\max} = \frac{g(g-1)}{2}$$

where $E_{\max}$ denotes the maximum number of edges for an undirected graph and g is the number of nodes (in our case *Departmental Programmes*). The density of a graph is simply the ratio of the edges actually present (L) to the maximum possible (Scott, 2000).

$$\Delta = \frac{2\boldsymbol{L}}{\boldsymbol{g(g-1)}} \quad 0 \leq \Delta \leq 1$$

Another insightful property of the network is the clustering or network transitivity (Watts & Strogatz, 1998; Watts, 1999). A network shows clustering if the probability of two nodes being connected by an edge is higher when the nodes in question have a common neighbor. One way of showing the existence of such a clustering in network data is to measure the fraction of "transitive triples" in a network (Wasserman & Faust, 1994), also called the clustering coefficient C (Watts & Strogatz, 1998). The Clustering Coefficient, C, is the average probability that two neighbors of a given node are also neighbors of each other and can be expressed as the proportion of triples that form a triangle out of all the triples present in the network. As Eckmann and Moses (2002) showed there is a close relation between highly clustered regions of a network and the existence of communities. In this study this measure helps to determine to which level the *Departmental Programmes* cluster together in each of the fields.

$$C = \frac{3 \text{ x } \textit{Number of triangles}}{\textit{number of connected triples}} \qquad 0 \leq C \leq 1$$

❖ *K-Core*

The additional analysis of the network was made using a technique to identify regions between the nodes called k-core. A k-core is a subgraph in which each node is connected to at least a minimum fixed number (K) of the other nodes in the subgraph (Seiman, 1983). The k-core approach allows actors to join the group if they are connected to k members, regardless of how many other members they may not be connected to (Wasserman & Faust, 1994). This resulted in network analyses on the level of these fields.

Research theme network analysis

Our goal on this part of the study is to identify communities between the researchers working in the LUMC assigned to a research theme and through correlation techniques (assortative mixing) measure patterns in the network structure.

We analyzed the level of interaction (based on co-publications) of the researchers from a research theme. To test the methodology we focused on just one of the research themes mentioned above: Neurosciences. Since our time period of analysis is 2002-2006 and the research themes were created in 2006, we are going to be able to measure whether the

neurosciences research theme which was created had already a level of synergy between researchers from different departments.

First we identify the communities and groups based on the co-publication networks and on an algorithm developed by Girvan and Newman (Girvan & Newman 2002; Newman 2004; Newman & Girvan 2004) based on the edge betweenness. Using the same idea of the node betweenness developed by Freeman (1977), the edge betweenness of an edge measures the times an edge is used in the shortest paths that connect two other nodes from the network. The edges that connect highly clustered communities have a higher betweenness so cutting these edges should separate communities. So the algorithm finds divisions of networks into closely knit groups by looking for the edges that connect groups (Lusseau & Newman, 2004).

Furthermore it is possible to measure whether the structure of the network is not randomly determined. This phenomenon is known as assortative mixing in networks (Newman, 2002; Newman, 2003; Newman & Park, 2003) in which the probability of two nodes being connected by an edge depends on some properties of those nodes. Assortative mixing on the basis of a scalar characteristic such as node degree is known as degree correlation. This measure determines whether there is preferential attachment between high-degree nodes and low-degree nodes, or whether there is preferential attachment between low and high degree nodes, referred to as disassortative mixing. Newman (2003) shows that it is possible to compute the degree correlation coefficient simply by calculating the Pearson correlation coefficient of the degrees at either ends of a link. This calculation should give a positive number for assortatively mixed networks and negative for disassortative ones. In terms of the network structure this will mean that a positive coefficient shows a core-periphery structure. The nodes with high degree are attracted with one another and as such form a core highly interconnected surrounded by a periphery of lower-degree nodes, on the other hand, negative coefficients cause the high-degree nodes to be scattered more broadly over the network.

3.4 Results

Departmental Programmes Analysis

As it was mentioned above, the overall research profile of the LUMC was used to create insight into the strength and weakness of the hospital. The standard impact of different main fields (more than 3%) of the research

profile of LUMC was measured through the field-normalized indicator CPP/FCSm.

Figure 1 shows the output and impact per field for the ten most prominent fields (accounting for at least 3% of the total of LUMC output in 1997-2006). 'Cardiovascular' is the most important field, including about 7% of the total output. Other three important fields are 'Hematology' with around 6% of the total output and 'Cardiac & Cardiovascular Systems' and 'Immunology' with almost 6%. The other important fields accounting between 5% and 3% of the total output are 'Endocrinology & Metabolism', 'Radiology', 'Nuclear Medicine & Medical Imaging', 'Biochemistry & Molecular Biology', 'Genetics & Heredity', 'Rheumatology', and 'Medicine, General & Internal'. The impact of these fields is in all cases above world average or world average (CPP/FCSm > 1).

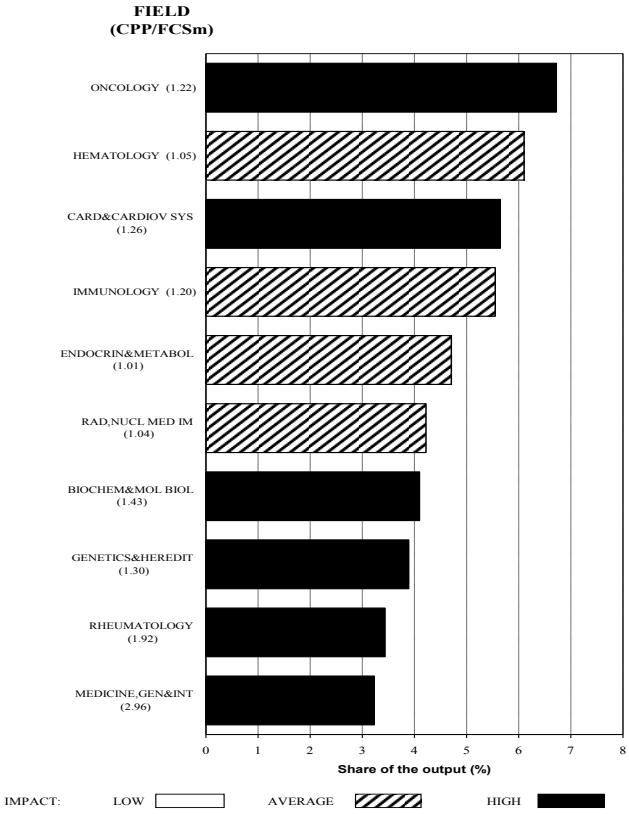

Figure 1. LUMC Research Profile, 1997-2006.

Using each of the inverse research *Programmes* profiles per research field, an analysis was conducted on the scientific cooperation relationships among the *Departmental Programmes*. Each of the 10 collaboration networks is then a set of research *Programmes*. The collaboration network analysis is focused on the structures per field within which the *Departmental Programmes* are embedded. Table 2 shows the result of the analysis where the above explained characteristics and properties of the networks are shown.

Table 2. Summary results: *Departmental Programmes* Network Overview

ISI Fields	*Depart Program*	Density	Clust Coeff	Main Core
Oncology	*56*	0.1305	0.371	7
Hematology	*53*	0.1183	0.308	5
Cardiac and Cardiovascular Sytems	*31*	0.1828	0.416	5
Immunology	*59*	0.1081	0.298	5
Endocrinology & Metabolism	*49*	0.1607	0.383	6
Radiology, Nuclear Medicine & Medical Imaging	*42*	0.1173	0.323	5
Biochemistry & Molecular Biology	*45*	0.1212	0.336	6
Genetics & Heredity	*47*	0.1045	0.306	5
Rheumatology	*35*	0.1277	0.333	4
Medicine, General & Internal	*40*	0.0962	0.427	4

The number of *Departmental Programmes* involved in each field gives already an approximation of the level of multidisciplinarity of the field. Fields like 'Immunology', 'Oncology' and 'Hematology' have many *Programmes* involved.

The 'Density' measures the general level of linkage between the *Programmes* and the clustering coefficient the level at which the network is clustered, which means that the network contains plocal communities in which a higher number of nodes create closely knit groups characterized by a relative high density of ties. These two indicators are complementary, so if the network has a high density and a high clustering coefficient (i.e. 'Cardiac and Cardiovascular Systems') this means that a group of *Programmes* tends to collaborate and this can provide insight for future merging and reorganization of the *Programmes*.

The last column shows the Highest K-Core. As it was mentioned before a k-core is a sub-graph in which each programme is connected to at least a minimum fixed number (K) of the other nodes in the sub-graph. Table 2

shows the highest core for each of the fields. Following with the 'Cardiac and Cardiovascular Systems' field, the highest core is a 5. This means that there is on the collaboration network of 31 Departmental Programmes a core of programmes which collaborates with at least five of the programmes from this core. In this case the 5-core contains 15 Departmental Programmes.

Research Theme Network analysis

We analyzed the level of interaction (based on co-publications) of the researchers from one of the research themes, Neurosciences. Our time period of analysis was 2002-2006. The principal researchers are highlighted with a red underlined on the researcher name.

Our goal was to identify the functional groups between the researchers assigned to a research theme. Figure 2 shows the groups identified by Girvan & Newman algorithm. Nodes coloring indicates group membership. The researchers part of the same group has the same color node. Overall the algorithm has identified six groups. The experts involved in the study commented that these groups fit quite well with the organizational groups attached to the principal researchers (highlighted in Figure 2) in the Neurosciences research theme. This is confirmed by the assortative mixing coefficient that has a negative value of -0.2154. The project leaders, having a lot of co-authored papers (high degree values) are spread on the network, collaborating mainly with the members of their groups.

The research themes were created in 2006 with the objective of promoting synergies between researchers from different departments who jointly study a single topic. As the analysis shows, the groups identified by the Newman and Girvan Algorithm fits quite well the real different research groups that will take part in the Neuroscience Research Theme. This means that in terms of collaboration activity for the period 2002-2006 the future members of the Neuroscience research theme have not yet started to collaborate, or at least not in terms of scientific production. The question would be how this collaboration network is at this moment. The creation and support of this research theme is having any effect in the collaboration activity between researchers coming from different disciplines?

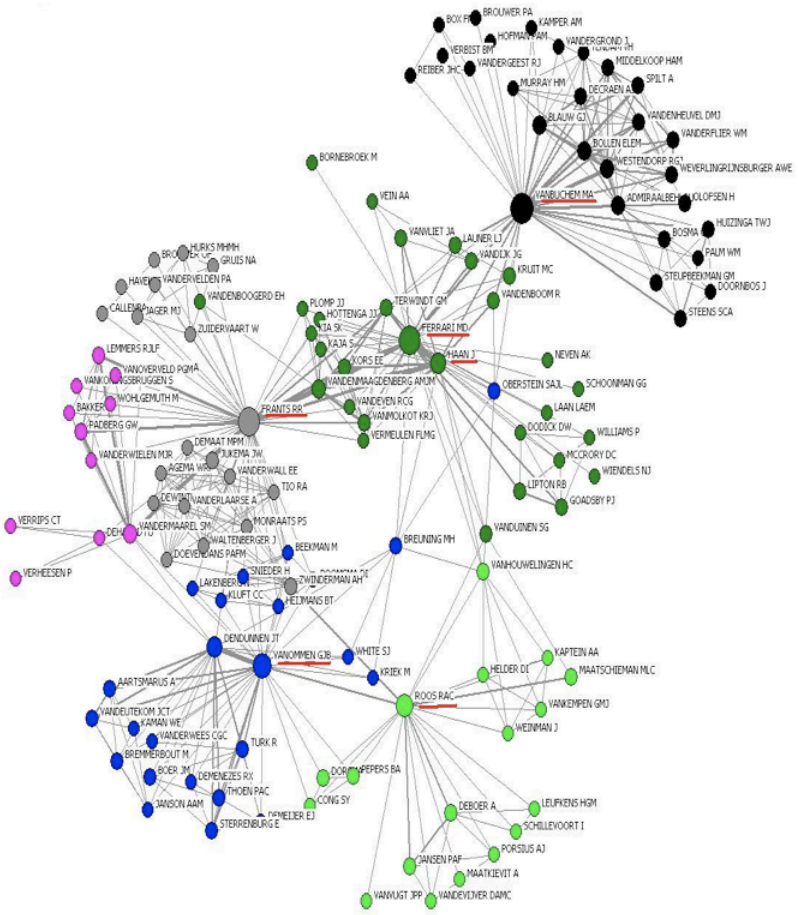

Figure 2. Collaboration Network: LUMC Neuroscience Research Theme

3.5 Conclusion

Translational research in a University Hospital is deeply embedded within daily work activities, it is not limited to a specific hierarchical or technical subset but highly distributed across the entire organization, this is why a proper management is very important to facilitate the research activities.

In the last years we have observed considerable advances in the development of methods for finding communities within networks, with

an enormous number of different techniques under development (Newman, 2008). The present study shows how bibliometric analysis can benefit from these developments and complement them. The combination of bibliometric indicators and network analysis can help the research managers of such as organizations to understand the way the organization behaves in order to create as strong as possible research clusters.

References

Barabási, A.L., Jeong, H., Ravas, E., Neda, Z.., Schuber, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations, Physica A, 311 : 590–614.

Börner, K., Chen C.M., and Boyeck K.W. (2003). Visualizing Knowledge Domains Annual Review of Information Science and Technology, 37, 179-255.

Bourke, P. and Butler, L. (1998). Institutions and the map of science: Matching university departments and fields of research. Research Policy, 26, 711-718.

Calero, C., Buter, R., Cabello, C., et al. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. Scientometrics, 66(2), 365–376.

De Nooy, W., Mrvar, A., and Batagelj, V. (2005). Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press.

Eckmann, J.P. and Moses, E. (2006). Curvature of co-links uncovers hidden thematic layers in the World Wide Weg. Proceedings of the National Academic of Sciences of the USA 99, 5825-5825.

Fontanarosa, P.B. and DeAngelis, C.D. (2002). Basic science and translational reseearch in JAMA. JAMA, 287 (13), 1728.

Freeman, L.C. (1977). A set of measures of centrality based upon betweenness. Sociometry, 40, 35-41.

Gibbons, M., Limoges, C., Nowotny, A., Schwartzman, S., Scott P., and Trow, M. (1994). The New Production of Knowledge: The Dynamics of Science and research in Contemporary Societies, Sage, London.

Girvan, M. and Newman, M.E.J. (2002). Community structure in social and biological networks. Proceedings of the National Academic of Sciences of the USA 99, 7821-7826.

Lambiotte, R., and P. Panzarasa (2009). Communities, Knowledge Creation and Information Diffusion. In: Katy Börner, Andrea Scharnhorst, (Eds.), Journal of Informetrics, Special Issue on the Science of Science: Conceptualizations and Models of Science, 3(3), 180–190.

Lusseau, D. and Newman, M.E.J. (2004). Identifying the role that animals play in their social networks. Proceedings of the Royal Society of London B 271 S6: S477-S481.

Newman, M.E.J. (2001). The structureof scientific collaboration networks, Proceedings of the National Academic of Sciences of the USA, 98 : 404–409.

Newman, M.E.J. (2002). Assortative mixing in networks. Physical. Review Letters 89, article no. 208701.

Newman, M.E.J. (2003). Mixing patterns in networks. Physical Review E 67, article no. 026126.

Newman, M.E.J. (2004). Detecting community structure in networks. The European Physical Journal B, 38(2):321–330, 2004.

Newman, M.E.J. (2008). The physics of networks. Physics Today, November 2008.

Newman, M.E.J. and Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E 69, article no. 026113.

Newman, M.E.J. and Park, J. (2003) Why social networks are different from other types of networks. Physical Review E 68, article no. 036122.

Moed, H.F., De Bruin, R.E., and Van Leeuwen, T.N. (1995). New bibliometric tools for the assessment of National Research Performance—Database description, overview of indicators and first applications, Scientometrics, 33 (3), 381–422.

Morillo, F., Bordons, M., and Gómez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. Scientometrics, 51, 203-222.

Pierce, S.J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. Journal of the American Society for Information Science and Technology, 50 (3), 271-279.

Qin, J., Lancaster, F. W., and Allen, B. (1997). Levels and types of collaboration in interdisciplinary research. Journal of the American Society for Information Science, 48(10), 893–916.

Rinia, E. (2000). Scientometric studies and their role in research policy of two research councils in the Netherlands, Scientometrics, 47 (2), 363–378.

Rinia, E.J., Van Leeuwen, T.N., Bruins, E.E.W., Van Vuren, H.G., & Van Raan, A.F.J. (2002). Measuring knowledge transfer between fields of science. Scientometrics, 54, 347-362.

Rinia, E.J., van Leeuwen, T.N.,Vuren, H.G., and van Raan, A.F.J., (2001). Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research. Research Policy 30, 357–361.

Rodriguez, M.A., and Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. Journal of Informetrics, 2(3), 195–201.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. Scientometrics, 59(3), 425-465.

Scott, J. (2000). Social Network Analysis: A Handbook. Newbury Park, CA: Sage Publications.

Seglen, P.O., and Aksnes, D.W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research, Scientometrics, 49: 125–143.

Seiman, S. (1983), Network structure and minimum degree, Social Networks, 5, 269-287.

Steele, T.W., and Stier J.C. (2000), The impact of interdisciplinary research in the environmental sciences: A forestry case study. Journal of the American Society for Information Science 51(5), 476-484.

Van Eck, N.J. and Waltman, L. (2007). Bibliometric mapping of the computational Intelligence Field. International Journal of Uncertainty, 15(5), 625-645.

Van Raan, A.F.J., and van Leeuwen, T.N. (2002). Assessment of the scientific basis of interdisciplinary, applied research. Application of bibliometric methods in Nutrition and Food Research. Research Policy, 31, 611-632.

Wasserman, S., and Faust, K. (1994). Social Network Analysis, Cambridge University Press, Cambridge.

Watts, D.J. (1999). Small Worlds: The Dynamics of Networks between Order and Randomness, Princeton University Press, Princeton.

Watts, D.J., and Strogatz, S. (1998). Collective dynamics of small-world networks. Nature.393:440-442.

Woolf, S.H. (2008). The meaning of Translational Research and why it matters. JAMA, 299 (2), 211-213.

Appendix

Table A. Departmental Programmes

Anesthesiology	
10100	No programme related research
10101	Perioperative Medicine: Efficacy, Safety and Outcome
Surgery	
10200	Traumatologie
10201	Vascular Surgery
10202	Surgical oncology
10203	Transplant surgery
Orthopedic Surgery	
10400	No programme related research
10401	Study of the normal and pathological locomotory system
10402	Diagnosis and treatment of bone and soft tissue tumours
Rehabilitation Medicine	
10500	No programme related research
10501	Pathofysiological analysis of movement disorders in relation to function
Thoracic Surgery	
10600	No programme related research
Urology	
10700	No programme related research
10701	Prostatic carcinoma
10702	Neuro-urology: functional disorders in male and female urogenital tract
Medical Decision Making	
10800	No programme related research
10801	Analysis and support of clinical decision making
Fysiotherapie	
10900	Fysiotherapy
Endocrinology	
20100	No programme related research
20101	Bone and mineral research
20102	Diabetes mellitus: pathophysiological changes and therapy
General Internal Medicine	
20201	The pathogenesis, clinical presentation and therapy of arterial and venous vascular disorders
Cardiology	
20300	No programme related research
20301	Vascular Biology and Intervention
20302	Cardiac Dysfunction and Arrhythmias
Pulmonology	
20400	No programme related research
20401	Pathogenesis and treatment of emphysema, other chronic obstructive pulmonary diseases and neoplasms of the lung
20402	Pathogenesis and treatment of asthma
Gastroenterology and Hepatology	
20500	No programme related research
20501	Cellular mechanisms in basic and clinical gastroenterology and hepatology
Nephrology	
20600	No programme related research
20601	Kidney and pancreas transplantation
20602	Vascular nephrology
Reumatology	
20700	No programme related research
20701	Pathophysiology and treatment of rheumatic diseases
Gerontology and Geriatrics	

20800	No programme related research
20801	Pathophysiology, epidemiology and therapy of ageing
Radiology	
20900	Onderzoeksprogramma's worden gereviseerd
Epidemiology	
21000	No programme related research
21001	Clinical epidemiology
Gynecology	
30100	No programme related research
30101	Cervix cancer
30102	Technology assessment of reproductive medicine
Obstetrics	
30200	No programme related research
30201	Research into fetal development and medicine
Dermatology and Venereology	
30400	No programme related research
30401	Dermatology-oncology
Otorhinolaryngology	
30500	No programme related research
30501	Disorders of the head and neck
Neurosurgery	
30600	No programme related research
30601	Assessment of spine and nerve surgeries
Neurology	
30700	No programme related research
30701	Pathophysiology of paroxysmal and chronic degenerative progressive disorder of the central and periferal nervous system
Ophthalmology	
30800	No programme related research
30801	Ophthalmic research
Pathology	
30900	No programme related research
30901	Immunopathology of vascular and renal diseases and of organ and celltransplantations
30902	Molecular tumour pathology - and tumour genetics
30903	Tumour immunology
Psychiatry (adults)	
31000	No programme related research
31001	Mood, anxiety and somatoform disorders and the HPA-axis (MASH)
Medical Psychology	
31100	Medical Psychology
Public Health	
31200	No programme related research
31201	Geriatrics in primary care
Paedriatics	
31300	No programme related research
31301	Stem cell transplantation and immunomodulation
31302	Epidemiology in Pediatrics and Child Health
31303	Development
CURIUM	
31400	No programme related research
31401	New methods for child psychiatric diagnosis and treatment outcome evaluation
Hematology	
40100	No programme related research
40101	Trombosis and Hemostasis
40103	Bone marrow failure
Immunohematology and Blood Transfusion	

40200	No programme related research
40202	Tumorimmunology
40203	Transplantation and autoimmunity
40204	Stemcel biology

Infectious Diseases

40300	No programme related research
40301	Antimicrobial treatment and prevention of infections
40302	Immunogenetics and cellular immunology of bacterial infectious diseases

Clinical Oncology

40400	No programme related research
40401	Experimental cancer immunology and therapy
40402	Biological, physical and clinical aspects of cancer treatment with ionising radiation
40403	Experimentele farmacotherapie

Clinical Pharmacy and Toxicology

40500	No programme related research
40501	Heterogeneity of drug efficacy and toxicity in relation to individual pharmacokinetics, pharmacodynamics and pharmacogenetics

Medical Microbiology

40600	No programme related research
40601	Molecular basis of virus replication, viral pathogenesis and antiviral strategies
40602	Molecular basis of bacterial pathogenesis, virulence factors and antibiotic resistance

Centraal Klinisch Chemisch Laboratorium

40700	Centraal Klinisch Chemisch Laboratorium Human Genetics
50100	No programme related research
50101	Mechanisms of disease, diagnostics and therapy
50102	Tumourgenetics and immunogenetics
50103	Genomics, epigenetics, population genetics and bioinformatics

Anatomy and Embryology

50200	No programme related research
50201	Molecular cardiovascular developmental biology

Molecular Cell Biology

50300	No programme related research
50301	Signal transduction in aging related diseases
50302	Gene regulation and cell differentiation
50303	Neurosciences in Drosophila and rodents; from genes to neuronal networks
50304	Microscopic imaging and technology

Parasitology

50400	No programme related research
50401	Host-parasite interactions with emphasis on immunology, molecular biologie, glycobiology and epidemiology of parasitic infections

Toxicogenetics

50500	No programme related research
50501	DNA repair mechanisms
50502	Replication associated mutagenesis
50503	Toxicogenomics and risk assessment

Medical Statistics and Bio Informatics

50600	COMICZ
50601	Development and application of statistical models for medical scientific research
50602	Molecular Epidemiology

Neuro-pharmacology

50700	No programme related research
50701	Stress hormones and brain function

Clinical Genetics

50800	No programme related research
50801	Genetics of disease, diagnosis and treatment
50802	Hereditary cancer genetics
50803	Genomics, population genetics and bioinformatics

4 Research cooperation within the bio-pharmaceutical industry: Network analyses of co-publications within and between firms

Clara Calero, Thed N. van Leeuwen, and Robert J.W. Tijssen
(published in Scientometrics 71(1): 87-99, 2007)

4.1 Introduction

Many of the largest bio-pharmaceutical firms face spend approximately 15% of their sales on R&D; in some cases their annual R&D costs amount to billions of euros. The globalization of markets, the regionalization of technical and scientific knowledge, scientific progress in the biomedical sciences, and the complexity of drug discovery processes, are forcing these companies to disperse their R&D organization and engage increasingly in R&D partnerships to access all the required knowledge and technologies. At the same time, as Howells (1990) points out, modern information and communication technologies serve to connect disseminated R&D activities and thus made distributed R&D organization possible. Because the bio-pharmaceuticals sector is often leading the way in this process of internationalization in their continual search for applicable knowledge and first-rate partners for their drug discover research, we are going to focus our attention on the large science-based multinational enterprises (MNEs) that are active in the bio-pharmaceuticals sector and produce relatively large numbers of research articles - either with partners within the MNE and/or with external partners within the private sector.

During the last two decades, the bio-pharmaceuticals industry has shifted its basic research operations from trial-and-error drug discovery approaches to a more science-based deductive method of searching for new target receptors and molecules that inhibit the target (e.g. Arora and Gambardella, 1994; Henderson and Cockburn, 1994; Nightingale, 2000). As a result, the ties between biotechnology companies and bio-pharmaceuticals companies have become close, and new organisational forms emerged to conduct basic bio-pharmaceutical research.

Traditionally a MNE's most strategic 'core' research activities were concentrated in a central R&D laboratory which was usually located in the home country. Nowadays, these elaborate organizational structures to enable research collaborations are determined and influenced by a wide range of factors, including the company's internal distribution and allocation of R&D resources (Gassmann and Van Zedtwitz, 1999; 2002), access to locally based technological expertise (Cantwell and Janne, 2000), the role of local or national governments in partnership promoting initiatives, as well as business strategies impacting on the propensity towards cooperation; outsourcing of research, or engaging in both horizontal (within-MNE) or vertical (external) research collaboration.

Clearly there are immense methodological problems in systematically analysing the organizational and geographical characteristics of R&D

activities by MNEs. Measuring and comparing their research cooperation networks is notoriously difficult. Many of these measurement problems relate to the scale, and levels of importance, of research cooperation, and the ways in which the objectives of research cooperation and networks can change over time. A solution for this dilemma is the application of empirical evidence extracted from the contents of scientific and technical articles that are authored industrial researchers and published in the peer-reviewed international scientific and technical journals. Although companies may publish for a large variety of reasons (Tijssen, 2004), one of which is to leverage results of their research as an interface to the global research community (Hicks, 1995), in most cases these articles reflect knowledge creation and knowledge transfer processes within corporate research labs.[1] The affiliate addresses of the author(s) listed on these research articles enable comparative analyses at the level of individual companies and their countries of location (e.g. Tijssen et al., 1996; Godin, 1996). Especially the big pharma companies publish sufficiently large annual quantities of research articles in the journal literature to warrant company-level (trend) analyses of research output related characteristics. A score of empirical studies have drawn on publication counts as an important indicator of research activity in the pharmaceutical industry (e.g. Koenig, 1983; Narin and Rozek, 1988; Gambardella, 1995; McMillan and Hamilton, 2000; Cockburn et al., 2000; and D'Este, 2005).

A significant share of industry's research articles list co-authors based at other affiliations within the same (parent) company, other companies, and/or public research organizations. This information source also enables aggregate-level quantitative information on patterns of research collaboration and related knowledge-spillovers.[2] We use these co-publications to examine structural factors that impact on research cooperation within and between companies. These research cooperation networks can be examined systemically by creating connectivity indicators based on these co-publications, thus showing relationships and linkages between the various actors and agents involved in joint scientific research. The network analyses of co-publication linkages indicate

[1] This source of printed 'codified knowledge' not only reflects "discovery" research done with the labs of the bio-pharmaceutics companies, but also related experience-based 'tacit' knowledge and the related skills base (e.g. Mowery et al., 1996).

[2] Co-authoring scientific publications is one of the clearest links to informal networking that can be made. These joint research papers reflect successful scientific co-operation and are likely to signify related knowledge flows and research networking activity between companies. Nevertheless, co-publication statistics and indicators should be handled with due care as a reliable source of conclusive empirical evidence on actual scientific cooperation (e.g. Katz and Martin, 1997).

structural differences between types of MNEs, which enabled us to develop a general typology of MNEs in terms of their patterns of research cooperation linkages.

4.2 Data collection and methodology

The research publications that were analysed for this study were extracted from CWTS's *Corporate Research Papers* (CRP) database, a subset of research articles published in international scientific and technical journals that are covered by the CWTS-licensed CD-Rom Edition of Thomson Scientific's *Citation Indexes*, in which at least one of the affiliate addresses of the authors refers to a private sector organization (see Tijssen, 2004).[3] A co-authored paper is fully credited to all firms listed in the author address information.[4,5] The bio-pharmaceuticals companies included in this study were selected according to their presence of at least one of their business unit or subsidiaries in Dunn & Bradstreet's "Who owns Whom" database (edition 2003), and the (parent) company's volume of (co-authored) research articles that were indexed in the CRP database. First, all business units/subsidiaries with Standard Industrial Code (SIC) code 2834 ("Drugs") were selected.[6] Then, all the (parent) companies of those business units/subsidiaries were selected that published at least five research articles during the years

[3] The CRP includes some 350,000 research papers published in 1996-2005 and (partially) assigned to the private sector. The coverage extends across all countries and fields of science and some 40,000 different main organizations are covered. Foreign branches and foreign subsidiaries of multinational companies are labelled with the consolidated name of the parent company. Companies that were added to the parent company through mergers and acquisitions were renamed to the current (ultimate) parent company to ensure backwards and forwards compatibility in trend analyses.

[4] Dividing a paper between the participating units (researchers, organizations, countries) is to some extent arbitrary - there is no fair method to determine how much money, effort, equipment and expertise each entity contributes the underlying research effort and writing the paper. Our basic assumption therefore is that each author, and associated corporate affiliate, made a non-negligible contribution.

[5] All co-publications are treated similarly in the statistical analyses, irrespective of the number or type of organisations (private or public sector) listed in the author address information. As a result, a co-publication listed two or more different (parent) companies may or may not include addresses referring to public sector organisations.

[6] Given the variety of SIC codes assigned to the different business units of the same (parent) company, many corporate affiliates were therefore allocated to several of the industrial sectors. The selection and matching procedure was carried out by CWTS in cooperation with CESPRI (Bocconi University, Milan, Italy).

1996-2001[7]. The companies that are represented in this set of publications were mostly (very) large pharmaceutical firms, especially the MNEs that invest heavily in their own research capabilities and sustain R&D labs that also perform original cutting-edge scientific research. One of the main characteristics of this set of papers is that 55.6% of the publications where produced by companies located in North America, including Canadian companies [8].

Each standardised name of the (parent) company was linked to the country of the company's location, i.e. the country of origin listed in the author affiliate address information in the research article. These company/country pairs enabled us to identify separate national affiliates of the (parent) companies (e.g. "Bayer AG/Germany" and "Bayer Corp/USA"), either the company headquarters in the home country or foreign subsidiaries. These company-country combinations are referred to as "corporate affiliates" hereafter. This breakdown by country enabled us to analyze and interpret co-publication data both in terms of intra-organizational collaboration (within the parent company "Bayer") as well as inter-organizational research partnerships (between different parent companies).

The last step to get the final core set of publications was to extract from the CRP database all the research papers that list at least two addresses referring to two selected corporate affiliates. The companies that took part on this set of publications were mostly (very) large pharmaceutical firms, especially MNEs. In total, there were 378 corporate affiliates. The resulting co-publication frequencies for each pair of affiliates were collected in a data array that was fed into the UCINET software package (Borgatti et al., 2002) that performs a network analysis providing statistics and graphical representations of the network structure.

[7] The data for this study were taken from the European project 'Network Indicators: Science, Tecnology and Innovation (STI-NET)". The STI-NET Project started on January 15th 2002 and was a 30 months project. The partners were CESPRI – Centre for Research on Innovation and Internationalisation Processes, Università Luigi Bocconi, Italy–, MERIT – Maastricht Economic Research Institute on Innovation and Technology, University of Limburg, The Netherlands- and the CWTS – The Centre for Science and Technology Studies, University of Leiden, The Netherlands– The aim of the research project was the identification, construction, and analytical use of network indicators in science, technology and innovation.

[8] For comparison, the US accounted for 31% of all research publications worldwide across all fields of science in 1998 and 2001 (EC, 2003), considerably less than its 55.6% share of corporate research papers in the bio-pharmaceutical industry.

4.3 Main Results

Distribution of co-publication partnerships by region

Using the geographical distribution of research partners listed on joint research papers, either within the same (ultimate) parent firm, or between firms, enabled us to estimate the degree of internationalization of corporate research cooperation. The breakdown by broad geographic region in Table 1 indicates that corporate research partnering within the pharmaceutical sector has become truly globalized at the end of the 20th century. Not only, do we also observe a particularly large propensity for tri-partite cooperation, with partners spread across three regions, which account for 15% of the co-publications, we also find 8% of the co-publications listing partners in four different regions. The majority of the relationships across three of four regions refer to connections between North America, EU15, and the Other European countries, the latter being largely a result of large MNEs based in Switzerland.

Table 1. Corporate co-publication partnerships by number of regions involved, 1996-2001

Location of research partners	Share of all co-publications (%)
Within regions – total	**46**
Within North America (NA)	24
Within Europe - EU15	12
Within Pacific Asia	10
Within Europe-Others	0
Two regions – total	**31**
NA + EU15	18
NA + Pacific Asia	5
NA + Europe-Others	2
EU15 + Europe-Others	3
EU15 + Pacific Asia	3
Three or four regions – total	**23**
NA + EU15 + Europe-Others	9
NA + EU15+Europe-Others + Pacific Asia	7
NA + EU15 + Pacific Asia	6
NA + Europe-Others + Pacific Asia	1

How can this degree of globalization and these geographical variations be explained? The above findings are obviously significantly affected by the geographically diversified science-based MNEs with R&D-labs scattered around the globe. Overall, we find a dominant role of North American companies in the trans-region research cooperation, which is in part due

to the attractive assets offered by US biotechnology companies in terms of their information sources, new approaches and advanced capabilities.[9]

This geographically diversified science-base is also influenced by the scattering of the research performing companies. Out of total of 378 affiliates: 42% were North American-based; 37% were located in EU15 countries; 18% in the Pacific Asia region (mainly Japan and Taiwan, and excluding Australia); 2% were based in 'Other European countries', principally in Switzerland, Norway, and Israel[10]; and only 1% were assigned to 'other countries' which refers to companies located in Australia.

The main determining factor for the North American surplus and deficits of co-publication outputs by the other regions clearly the size of the (science-related) industrial base of each region. In terms of sheer magnitude, we note the marked dominance of North American-based affiliates (US and Canada) as a consequence from the combination of several factors: (1) the size of US industry in the sectors, (2) the scale of their research activity, (3) their propensity for research cooperation, (4) their propensity to publish research findings in the open scientific literature. These factors are obviously difficult to unravel empirically at an aggregate level, let alone at the company level, a second explanatory determinant can be gleaned from further breaking down the partnerships geographically.

Table 2 exhibits the breakdown of co-publications by geographic location of research partners. First, we observe a strong tendency towards corporate research partnering within the region. The majority of the co-publication partners of each of the major regions (North America, Europe, and Pacific Asia) are located within the same region, which is to a large degree the logical consequence of proximity effects due to common (domestic) research systems, share language and culture, or the regional scope of business activities. However, we do find a relatively low share of within-region partnerships within the EU15 in comparison to North America, which suggests that the European pharmaceutical industry less affected by proximity effects. One of the reasons for this difference is the US-orientation of the large Swiss bio-pharmaceutical MNEs (for example *Novartis*, which is further discussed below). Interestingly, these non-EU15 companies prefer North America partners even more than EU15 counterparts. Another possible explanation for

[9] One of the major strengths of the US bio-pharmaceuticals industry lies in its specialization in bio/gene technologies for drug discovery in the fields of immunology and oncology.
[10] Israel was considered a European country in this study.

Europe's extra-regional orientation relates to the presence of affiliates of non-European MNEs within Europe; these foreign affiliates tend to engage in intra-MNE cooperation with other affiliates located outside Europe (as illustrated in the examples provided below).

Table 2. Within-region and between-region research collaboration between corporate affiliates 1996-2001 (row percentages)

Location of affiliate	% within-region partnerships	% partners in other regions				
		NA	EU15	Europe-Other	Pacific Asia	Other
North America (NA)	75		16	4	2	3
Europe - EU15	52	37		8	3	0
Europe - Other	12	48	37		4	0
Pacific Asia	66	19	12	3	0	0

When analyzing the intra-MNE organisational dimension of research networks, size does seem to matter. Considering their large share of the corporate research literature worldwide, the US-based affiliations seem to have relatively many domestic partners. The domestic orientation of the USA is in part explained by the scale and diversity of the US corporate sector, which comprises of an almost self-contained R&D base. We find a significant overrepresentation of the EU15-based affiliates in research partnerships, which is in large part due to the European affiliates of US companies. Remarkable is the relatively small share of partnerships between North America and the companies in Pacific Asia. This deviant result might reflect cultural differences, where Western companies tend to adopt relatively "open" R&D models with international cooperative structures whereas MNEs with headquarters based in Japan and South Korea prefer more "closed" structures that favour cooperation across close geographic proximities. In contrast, the EU-15 based MNEs are much less regionally focused. Next, we observe remarkably low intra-regional cooperation propensities within the affiliates based in the Other European countries, which is no doubt due to the small size of the domestic industrial base - hence, few (potential) research partners - and partially a result of the industrial sector structure which is dominated by a few large multinational affiliates with branches and R&D labs worldwide (e.g. Swiss-based pharmaceutics companies).

Research collaboration network patterns

Only 35 companies (accounting for 129 corporate affiliates) are in fact MNEs exhibiting co-publication links between their various affiliations (i.e. company headquarters and subsidiaries/branches in different countries). The remainder comprises multinational enterprises with only one research facility or R&D laboratory that is producing research papers, and national enterprises with no foreign affiliations. Starting from the arrays of co-authorship frequencies, the co-publication linkages between the MNE's headquarters and its foreign subsidiaries are identified and the various nodes (circles), each representing corporate affiliates, are connected. Each link between the nodes indicates one or more co-authored research publications. Co-publication links to affiliates of other companies are represented by squares.

Focusing on the structural features of the co-publication patterns of those 31 'networked' MNEs, we identify three general types of co-publication networks characterizing the relationships between headquarters and subsidiaries of the same company: (1) centralized networks, (2) decentralized networks, and (3) gateway networks.

Centralized corporate research network. This particular kind of network consists of co-publications involving the company headquarters (i.e. central R&D laboratory) and its affiliates, as well as their co-publishing activity with other companies that tend to concentrate at the company headquarters. This set includes 14 MNEs. Figure 1 displays an illustrative example of this type of centralized network for the case of *Bayer*. All of Bayer's subsidiaries co-published with the headquarters in Germany, as well as with Bayer labs in the USA. The headquarters, USA and Great Britain all have research collaborations with other firms based in the EU15, in other European countries and in North America. Bayer/Japan extends this network with additional collaborations with companies in its own region.

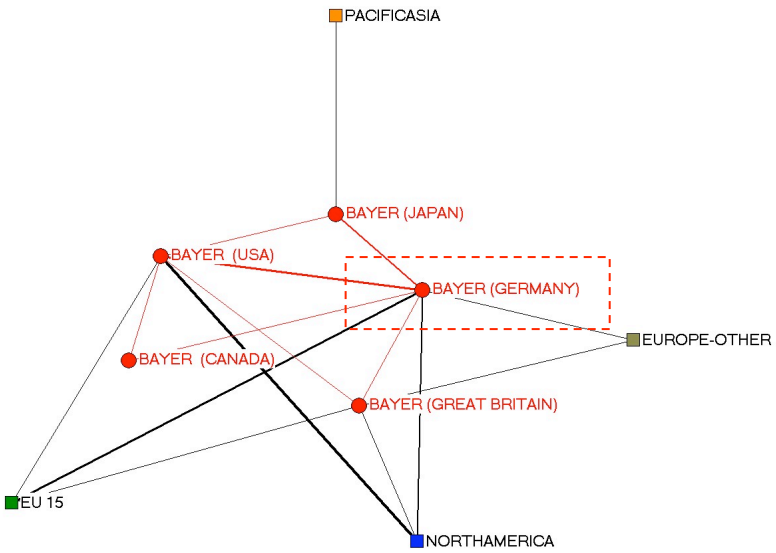

Figure 1. Centralized corporate research network: Bayer

Circles.Country of location of corporate affiliates belonging to the same multinational enterprise.
Squares.Region of location of external partners in research co-publications.
(The thickness of the connecting line indicates the quantity of co-publications.)

Decentralized corporate research networks. This category comprises of networks that are characterized by the lack of (strong) links between the company headquarters (i.e. central research laboratory) and subsidiaries that do not co-publish. Rather, each research laboratory co-publishes with other companies located in the different regions of the globe. The pattern reflects a corporate R&D strategy with geographically dispersed and (semi) autonomous research laboratories. About 14 of the 31 MNEs can be attributed to this category. Figure 2 shows then network of *Novartis*. Novartis' headquarters are located in Switzerland with several subsidiaries scattered across the globe. Novartis Switzerland is connected with all regions, with an especially strong link to Novartis' operations in the EU-15 and the North America region. We can see how each of the subsidiaries have their own pattern of collaboration, sometimes focused on EU-15 and North America region, as the case of Novartis' research in Netherlands, Italy, France, while others focus more on Pacific Asia and North America region, like in the case of Novartis research activities in Japan.

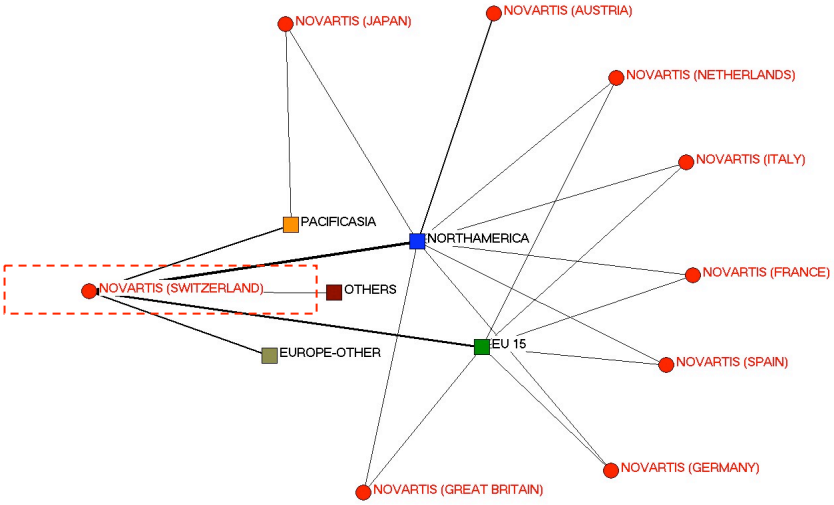

Figure 2. Decentralized corporate research network: Novartis

Circles.Country of location of corporate affiliates belonging to the same multinational enterprise.
Squares.Region of location of external partners in research co-publications.
(The thickness of the connecting line indicates the quantity of co-publications.)

Gateway corporate research network. This kind of network is characterized by researchers and scientists at the company's subsidiaries co-publishing exclusively with their colleagues based at headquarters/central R&D facility, while the latter also co-publish with external partners. Only a small minority of the 31 companies appears to operate through these gateway networks. This rare kind of network is characterized by researchers and scientists at the company's subsidiaries co-publishing exclusively with their colleagues based at headquarters/central R&D facility, while the latter also co-publish with external partners. Figure 3 displays the example of *Baxter Healthcare*, a US company with two subsidiaries that run research facilities – Baxter Germany and Baxter Belgium. Each subsidiary co-publishes with their headquarters, which in turn co-publishes also with other companies located elsewhere, North America and EU-15.

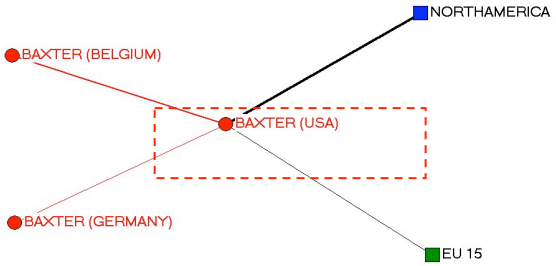

Figure 3. Gateway corporate research network: Baxter

Circles.Country of location of corporate affiliates belonging to the same multinational enterprise.
Squares.- Region of location of external partners in research co-publications.
(The thickness of the connecting line indicates the quantity of co-publications.)

4.4 Discussion and concluding remarks

This empirical study set out to explore the analytical potential of corporate research articles as a source of empirical information for describing structural patterns of research joint ventures (RJVs) within the bio-pharmaceutical industry worldwide, and to produce quantitative data on those research cooperation relationships at the level of countries and major bio-pharmaceutical firms. Given the overwhelming significance of basic research in the bio-pharmaceutical industry, and the large quantity of corporate research papers produced each year, we believe than these publications reflect key characteristics of research cooperation patterns within the industry. The pivotal position of the USA in the bio-pharmaceuticals research output, and the associated global research network, is not surprising in view of the US dominance in the bio-pharmaceutical sector. More interesting is the particularly strong link we observe between US companies and their research partners in EU-15 countries – either with overseas affiliations of these US companies or external partners. At the firm level, we discerned two types of links: (1) those amongst corporate affiliations belonging to the same (parent)

company, and (2) linkages between affiliations of different (parent) companies. The outcome revealed interesting empirical information both with respect to the organizational features of corporate research partnerships within and between companies, as well as geographical dispersion of these partnerships. The company-level breakdown of these cooperation patterns also reveals a variety of intra-firm and extra-firm research linkages, from which three main types of corporate research networks can be derived in terms of the intra-firm distribution of research partnerships: (a) centralized networks, (b) decentralized networks, and (c) gateway networks. The first two types seem to be by far the most common ones.

It stands to reason that the various types of within-firm linkages are driven by different corporate "logic" governed by the prevailing R&D objectives and constraints, intellectual property rights and knowledge appropriation regimes, and research cooperation motives. Moreover, some of the large pharmaceutical companies nowadays adopt "open" innovation structures, where R&D cooperation and networking both within and outside the company become increasingly integrated, especially between 'big pharma' companies and smaller biotechnology companies (a recent example is the relationship between Roch, the Swiss company, and the UK biotech firm Antisoma). The joint research publications emerging from these pharma-biotech RJVs have not been included in this study owing to the relatively low numbers of papers.

It is also important to note that many of these research partnerships, and the corresponding co-publications, may also include partners from public sector research organisations and universities (Tijssen, 2004). These contributions were not included in the network analysis presented in this paper and are left to future research. Nonetheless, the presence of public sector researchers in 'corporate' RJVs raises questions about the nature of the research links between the various corporate affiliations; to what degree these were predominantly curiosity-driven 'academic' partnerships, industry-driven 'application-oriented' partnerships, or a mixture of both? Each type of collaboration is likely to operate according to their own rationale, with different sets of (ultimate) goals and deliverables. Generally speaking, these joint research articles should be viewed as reflecting research cooperation at the work floor level. As such, they are more likely to describe 'informal' research linkages and networking processes between individual researchers, rather than representing the key results of 'formalised' and targeted alliances between R&D departments or research teams within firms. As a consequence, the structural characteristics of co-publication networks, as depicted in our graphs, are therefore not likely to correspond or correlate

with senior management's view of its networking activity in the same way as linkages based on corporate R&D alliances. We would argue that a publications-based view of corporate research is in fact one of the strengths of our approach, in the sense that it helps external analysts get closer to the joint research products emerging from the day-to-day operations of scientists and technicians employed by the research labs of the bio-pharmaceutical companies.

To conclude, although our empirical data shed some light on corporate research partnering in the bio-pharmaceutical industry, especially within large science-intensive multinational enterprises, we still know little about the detailed and hard-to-observe mechanisms and organizational conditions giving rise to research articles produced by corporate researchers in collaboration with colleagues. Case studies of individual pharmaceutical companies, such as recent studies conducted by Criscuolo and Narula (2005) or Criscuolo (2005), constitute important next steps to help gain inside-information to unravel the country-specific, firm-specific and person-specific determinants that impact on the reasons for engaging in research cooperation and the propensity to produce the (joint) research articles we have analyzed in this study.

References

Arora, A., Gambardella, A. (1994), The changing technology of technological change: general and abstract knowledge and the division of the innovative labour. Research Policy, 23, 523-532.

Borgatti, S. P., Everett, M. G., Freeman, L. C. (2002), Ucinet 6 for Windows. Harvard: Analytic Technologies.

Brusconi, S., Criscuolo, P., Geuna, A. (2005), The knowledge bases of the world's largest bio-pharmaceuticals groups: what do the patent citations to non-patent literature reveal? Economics of Innovation and New Technology, 14, 395-415.

Cantwell, J., Janne, O. (2000), The role of multinational corporations and national states in the globalization of innovatory capacity: the european perspective. Technology Analysis and Strategic Management, 12, 243-262.

Cockburn, I.M., Henderson, R.M.,, Stern, S. (2000), Untangling the origins of competitive advantage. Strategic Management Journal, 21, 1123-1145.

Coombs, R., Georghiou, L. (2002), A new "Industrial Ecology". Science, 296, 471.

Criscuolo, P. (2005), On the road again: Researcher mobility inside the R&D network. Research Policy, 34,1350-1365.

Criscuolo P., Narula R. (2005), Using multi-hub structures for international R&D: organizational inertia and the challenges of implementation, Working paper.

D'Este P. (2005), How do firms' knowledge bases affect intra-industry heterogeneity? An analysis of the Spanish pharmaceutical industry. Research Policy, 34, 33-45.

EC. (2003), Third European Report on Science and Technology Indicators: towards a knowledge-based economy. Brussels: European Commission, Report EUR 200025.

Gambardella, A. (1995) Science and Innovation. The US Pharmaceutical Industry during the 1980s. Cambridge University Press, Cambridge.

Gassmann, O., von Zedtwitz, M. (1999), New concepts and trends in international R&D organization. Research Policy, 28, 231–250.

Godin, B. (1996), Research and the practice of publication in industries. Research Policy 25, 587–606.

Hagedoorn, J., Link, A., Vonortas, N. (2000), Research Partnerships. Research Policy, 29, 567-586.

Henderson, R., Cockburn, I. (1994), Measuring competence? Exploring firm effects in pharmaceutical research. Strategic Management Journal ,15 (Winter), 63-84.

Hicks, D.M. (1995), Published Papers, Tacit Competencies and Corporate Management of the Public/Private Character of Knowledge. Industrial and Corporate Change, 4, 401-424.

Howells, J. (1990), The location, organisation of research and development: new horizons. Research Policy, 19, 133–146.

Jaffe, A. (1989), Real Effects of Academic Research, American Economic Review, 79, 957-970.

Koenig, M.E.D. (1983), A bibliometric analysis of pharmaceutical research. Research Policy, 12 ,15-36.

Katz, J.S., Martin, B.R. (1997), What is research collaboration? Research Policy, 26, 1-18.

McMillan, G.S., Hamilton, R.D. (2000), Using bibliometrics to measure firm knowledge: an analysis of the US pharmaceutical industry. Technology Analysis and Strategic Management, 12, 465-475.

Mowery, D.C., Oxley, J.E., Silverman, B.S. (1996), Strategic alliances and interfirm knowledge transfer. Strategic Management Journal, 17, 77-91.

Narin, F., Rozek R.P. (1988), Bibliometric Analysis of bibliometric analysis of united-states pharmaceutical-industry research performance. Research Policy, 17, 139-154.

Nightingale, P. (2000), Economies of scale in experimentation: knowledge and technology in pharmaceutical R&D. Industrial and Corporate Change, 9, 315-359.

Pavitt, K. (1998), The social shaping of the national science base. Research Policy, 27, 793-805.

Tijssen, R.J.W., T. Van Leeuwen,, Korevaar J.C. (1996), Scientific publication activity of industry in the Netherlands. Research Evaluation, 6, 1-15.

Tijssen, R.J.W. (2004), Is the commercialisation of scientific research affecting the production of public knowledge? Global trends in the output of corporate research articles. Research Policy, 33, 709-733.

Von Zedtwitz, M., Gassmann, O. (2002), Market versus technology drive in R&D internationalisation: four different patterns of managing research and development. Research Policy, 31, 569–588.

5 Important factors when interpreting bibliometric rankings of world universities: an example from oncology

Clara Calero-Medina, Carmen López-Illescas, Martijn S. Visser, Henk F. Moed (published in Research Evaluation 17(1): 71-81, 2008)

5.1 Introduction

More and more attempts are made to identify top research universities from a global perspective as internationalization and globalization in academic research and teaching proceeds, Universities are more and more competing for research funds, research students and researchers in the global research area. Their reputation as research universities is a crucial factor in such a competitive system. Therefore, members of the international scientific community, officials responsible for institutional, national and supra-national science policies, and the wider public need 'objective', 'reliable' information about the research performance of universities.

Comparative analyses of the performance of universities at a national level, focusing on particular research fields or disciplines, have been carried out for many years. For instance, in 1995 the US National Research Council (NRC), the working arm of the National Academy of Science and the National Academy of Engineering, published a report presenting a quality rating of PhD programs at 274 US institutions in 41 fields, based on surveys sent to faculty (Goldberger et al, 1995). The NRC report also presented bibliometric indicators based on publication and citation data extracted from the ISI Citation Indexes, but these indicators were not used by the NRC for ranking purposes. Diamond and Graham (2000) further analysed the NRC data and concluded that "reputational ratings showed a strong positive correlation with citation densities", in the sense that the institutions appearing in the top of the former tended to be highly ranking on the latter as well. However, younger and smaller "challenging" institutions tended to have higher positions in the citation impact rankings than in the reputational rankings.

A recent phenomenon is the compilation of rankings of universities from a supra-national or global perspective. For instance, the European Commission published in the recent European Science Indicators Reports listings of European universities presenting their bibliometric scores. Global rankings of universities were published by the Jiao Tong University in Shanghai (SJTU, 2007) and by the Times Higher Education Supplement (THES, 2007). The SJTU rankings were to a large extent based upon bibliometric indicators, and partly upon counts of prizes and awards. In compiling the THES rankings, expert opinions collected from surveys constituted the most important indicator, while bibliometric indicators played a less important role. For a thorough review of these two rankings, the reader is referred to Van Raan (2005).

This paper presents bibliometric characteristics of the 386 most frequently publishing world universities and of a (partly overlapping) set of 529 European universities. Rather than showing a ranking itself, it presents a statistical analysis of ranking data, focusing on more general patterns in the data. It compares US universities with European institutions; countries with a strong concentration of academic research activities among universities with nations showing a more even distribution; a ranking of universities based on indicators calculated for all research fields combined with one compiled for a single field; general with specialised universities; and rankings based on a single indicator with maps combining social network analysis and a series of indicators. It highlights important factors that should be taken into account in the interpretation of rankings of research universities based on bibliometric indicators. Moreover, it illustrates policy-relevant research questions that may be addressed in secondary analyses of ranking data. In this way, this paper aims at contributing to a public information system on universities, particularly research universities, useful in research management and policy at the institutional and (supra-)national level, and for the wider public.

The paper provides a series of bibliometric indicators of the research performance of universities, derived from the Web of Science (WoS), published by Thomson Scientific. Research universities produce knowledge, contribute to the advancement of scientific-scholarly knowledge. These contributions are normally embodied in research articles, published in the open, serial literature and subjected to criticism of colleagues. A base assumption underlying a bibliometric approach is that one can learn about scientific activity and performance by analyzing the scientific literature (e.g. Garfield, 1964, 1979; Narin, 1976). In this paper three bibliometric indicators play a key role, measuring article production, disciplinary specialisation, and citation impact, respectively. A brief description of the methodology applied in this paper is given.

A first research question is: How does the citation impact of European universities relate to that of their US counterparts? A basic notion underlying the analyses presented under the heading ‘Comparison of European and US universities’ holds that the bibliometric outcomes of an individual university can only be interpreted properly when one takes into account the structure of the national academic system in which it is embedded, and the particular role of the university therein. The next sections distinguish two models for distributing ‘top’ research among a nation’s universities: a concentration model in which a limited number of big research universities carries out research at a top level in a wide range of disciplines, and a distributed model, in which top research is more

evenly distributed among universities, and a strong link between teaching and research is maintained. In the USA the concentration model is dominant, whereas in Europe many countries tend to show a more distributed model, although substantial differences exist among European countries.

These differences in the degree of concentration of a country's academic research activities among its universities are further analysed. In the set of European countries, the statistical relationship between a country's degree of concentration within the academic system and its overall performance measured in terms of citation impact is examined. The research question addressed is: Do European countries in which academic research activities are concentrated in a limited number of universities perform better than nations in which research is more evenly distributed among its academic institutions?

Rankings of world universities are normally based on indicators for an institution as a whole, combining all fields in which it is active. Universities tend to be active in a range of scientific-scholarly research fields, but their performance may vary from one field to another. This variability is invisible in an overall indicator such as the total number of published articles, or the normalised citation impact calculated for a university's total publication output. The next section addresses the following question: To what extent does a ranking of universities based on their bibliometric scores in a particular research field differ from that based on an overall indicator calculated for all fields combined? As an example, the field of oncology is analysed.

The distinction between general and specialised universities is highlighted, even though it is difficult to draw a sharp borderline between the two. An index is proposed to measure the degree of disciplinary specialisation in a university's research activities, applying a classification of published articles into 15 disciplines. A research question addressed is: How does the performance of general universities statistically relate to that of specialised universities? The section compares the citation impact of the two types of academic institutions, both at the level of an institution as a whole and at the level of individual disciplines.

Rankings are in a sense one-dimensional, as entities are ordered by descending score on one particular statistic, even if it is a compound measure based on a weighted series of indicators. Rankings disregard how the performance of one entity depends upon that of others. The next section deals with the question: what are the potentialities of using social

network analysis to display collaboration networks among universities? It presents preliminary outcomes of an analysis of the top 100 world universities in terms of number of published articles.

Finally, the conclusion indicates lines of future research, and proposes further steps towards the creation of a reliable information system of world universities, and its use in thorough empirical analyses of policy relevant issues.

5.2 General methodology

Assignment of articles to universities; accuracy

For European universities the Membership Directory of the European University Association was used as a starting point. Since this list did not include all European universities, it was expanded during the project. The data collection process aimed at defining the article output of European universities publishing at least 500 papers during the time period 1997–2004. For non-European universities the process identified the articles of the 200 most frequently publishing universities. Articles were assigned to universities on the basis of the information on the institutional affiliations of authors, included in the corporate address field. Two rounds were carried out.

In a first round, papers were selected with the name of a university (and its major departments) mentioned explicitly in the address. Name variations were taken into account. For instance, Ruprecht Karls University is a name variant of the University of Heidelberg, TUM of the Technical University München; and Université Paris 06 of Université Pierre et Marie Curie. For European universities, this round took into account all variations occurring five or more times. For non-European universities this threshold was set to 25.

In a second round, additional papers were selected from affiliated teaching hospitals on the basis of an author analysis. This round added to a particular university's article output selected in the first round papers from affiliated hospitals, published by authors who did not explicitly mention this university's name in their institutional affiliation, but who showed strong collaboration links with that university, as its name appeared in the address lists of at least half of their papers. In this way, for instance, a part of the papers containing the address Addenbrookes Hospital was assigned to University Cambridge, and a part of the papers

with the address Hospital La Pitié Salpetrière to University of Paris VI, and another part to University of Paris V.

Since the de-duplication and counting process of European universities took into account only name variants occurring five or more times, an overall accuracy rate for this group of universities is estimated to be about 95%. It is somewhat higher for universities with a large number of published articles than it is for universities with smaller publication volumes. For non-European universities it is around 90%. It is important to note that the data were not verified by representatives of institutions.

Universities analysed in this paper

This paper analyses two sets of universities. The first and most important one is the set of universities that published more than 5,000 articles in WoS journals during 1997–2004, or on average more than 625 papers per year during this time period. It contained 386 universities, and is denoted as the global or world set, containing world universities. In view of the collaboration among institutions, resulting in co-publications by scientists from two or more institutions, it would be more precise to state that the universities contributed at least one author to more than 625 papers per year. Technically, this number is denoted as an integer count. A second set of universities analysed in this paper is a set of 529 European universities publishing at least 500 articles during 1997–2004, or on average 65 articles per year. There is an overlap between the European and the global set: 172 European universities are included in both sets.

Indicators calculated

The indicators calculated in this paper are summarized in Table 1. The first indicator, denoted as article output, is defined as the number of articles published during a particular time period in journals processed for the WoS. Article types included in the counts are full articles, letters and reviews. Other types, such as editorials, discussion papers and meeting abstracts, are not included.

A disciplinary specialisation index for a particular university is based on Pratt's Index, calculated for a university's distribution of normalised publication activity across 15 disciplines. These disciplines are listed in Table 2. Pratt's Index ranges between 0 (no specialisation at all) and 1 (extremely strong specialization). For further details on this index the reader is referred to Moed (2006), Bookstein and Yitzhaki (1999) and Egghe and Rousseau (1990).

Normalised citation impact is defined as the average number of citations per article published from a university, relative to the world citation average in the subfields in which it is active. It is also denoted below as 'citation impact' or 'impact per paper'. A value of 1.0 indicates a citation impact equal to the world citation average. Details can be found in Moed et al, 1995) or in Van Raan (1996).

Table 1. Four bibliometric indicators calculated in this paper

Indicator	*What it measures*	*Technical description*
Article output	Scale of scientific activity (number of active scientists) and article productivity (number of articles per active scientist)	The number of research articles published in about 7,500 journals processed for the *WOS*
Disciplinary specialisation index	Are activities more or less evenly distributed among disciplines (as in general universities) or concentrated (as for instance in medical, agricultural or technical universities)?	Pratt Index: ranges between 0 (no specialisation at all) and 1 (extremely strong specialisation); assessed relative to the world distribution.
Normalised citation impact (also denoted as citation impact per paper)	Intellectual influence; prominence of research groups in their fields; their authoritativeness; visibility	Average number of citations per article published by a university, relative to the world citation average in the subfields in which it is active
Collaboration strength	The extent to which two universities collaborate as expressed in co-authorship	Number of co-publications between two universities, divided by the square root of the product of the number of papers published by each.

Table 2. Pearson correlation coefficients between a university's normalised citation impact in a discipline and its publication activity index in that discipline

Acronym	*Discipline*	*Univs*	*Pearson's R*
APC	Applied physics & chemistry	270	–0.02
BIOL-HU	Biological sciences primarily related to humans	310	+0.24 *
BIOL-AP	Biological sciences primarily related to animals and plants	194	–0.18
CHEM	Chemistry	301	–0.03
CLM	Clinical medicine	320	+0.23 *
ECON	Economics	23	–0.05
ENG	Engineering	227	–0.03
GEO	Geosciences	147	+0.15
A&H	Humanities & arts	40	+0.12
MATH	Mathematics	75	–0.11
MOLB	Molecular biology & biochemistry	270	+0.41*
SOC-MED	Other social sciences primarily related to medicine & health	81	+0.01
SOC	Other social sciences	70	–0.20
PHYS	Physics & astronomy	290	+0.17 *
PSY	Psychology & psychiatry	101	–0.07

Note:* Significant at p=0.01.

5.3 Comparison of European and US universities[1]

Figure 1 relates to the 386 universities publishing more than 5,000 papers during the time period 1997–2004. The horizontal axis gives the average number of articles published per year during this time period, and the vertical axis their normalized citation impact. Universities are categorized into three broad geographical regions: USA, Europe and all other countries.

Figure 1 shows that US universities are highly overrepresented in the top of the ranking based on normalised citation impact, and to a lesser extent, on the number of published articles per year. In fact, in the group of the 25 universities with the highest citation impact, all universities are from the USA, and in the group of 76 universities with a citation impact above 1.5, 67 (88 per cent) are located in the USA. Among the top 25

[1] Sections 5.3 and 5.4 are partly based upon: Visser, M.S., Calero-Medina, C. and Moed, H.F. (2007). Beyond rankings: The role of large research universities in the global scientific communication system. In: Torres-Salinas, D. and Moed, H.F. (eds.). Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, 25-27 June, 2007, Vol II, 761-765.

institutions with the highest number of published articles per year, 20 (80 per cent) are from the USA.
In the set of 386 world universities, 172 are located in Europe, and 122 in the USA. Table 3 gives for each geographical region the mean and quartiles of the distribution of normalised citation impact among universities. Table 3 shows that US universities tend to have a higher normalised citation impact than European academic institutions: 1.55 versus 1.11. The 75th percentile of the distribution for Europe is lower than the 25th percentile for the USA. The third column shows that the 172 European universities account for around 72 per cent of the total European university output. The 122 US universities published about 83 per cent of the total US university output. This percentage is higher than the 72 per cent obtained for Europe, and indicates that there is a stronger concentration of published articles among US universities than there is among European institutions, in agreement with earlier analyses published by Matia et al. (2005).

In order to further characterize differences among European and US universities, an institution's citation impact was analysed *per discipline*, using a classification of research articles into 15 disciplines, listed in Table 2. More detailed information can be obtained from Moed (2005: 189). For each institution the number of disciplines was determined in which it was 'world leader', i.e. ranked among the op 10 or top 25% according to the normalized citation impact in the set of 386 world universities. For each geographical region, the number and percentage of universities was determined that was world leader in at least one discipline, and for these institutions the average number of such 'top' disciplines per university was computed. These indicators were calculated for all universities in the set, and also for the 'very best' universities in their region, i.e. being among top 25% in their region on the basis of their overall normalised citation impact.

The results are presented in Table 4. The upper half of this table presents the outcomes when the concept of 'world leader' in a discipline is defined as being among the top 10% among all 386 world universities in that discipline. In the lower half, the criterion for being world leader is somewhat relaxed, and defined as belonging to the top 25% in a discipline. A key finding is that all the very best European universities are among the 25% best in the world in at least one discipline, and 65% of them even in the top 10% in a field, but that the number of disciplines in which they are world leader is on average substantially lower than that for top US universities.

In a recent report, Lambert and Butler (2006) analysed differences among continental European countries, the UK and the USA as regards the structure and research performance of their national academic systems. They mentioned several structural factors that in their view are responsible for what they term as 'mediocrity' of (particularly continental) European universities, including a lack of concentration of funds among institutions. The citation impact analysis presented in this section indeed revealed that the overall citation impact per paper of European universities tends to be lower than that of their US counterparts. One may question whether the term 'mediocre' is appropriate to qualify the position of European universities in the world rankings. But even if one adopts this qualification from Lambert and Butler, it needs emphasising that 'mediocrity' of a university does not necessarily imply that it is mediocre in *all* disciplines.

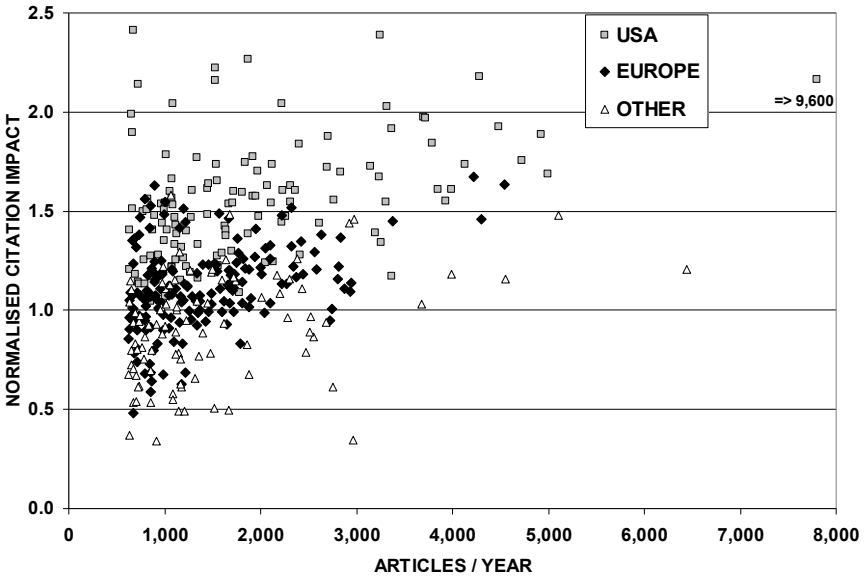

Figure 1. Number of published articles per year and normalised citation impact for world universities

Table 3. Distribution of citation impact among European and US universities

Region	*No Universities*	*% Papers from univs*	*Normalised citation impact distribution*			
			Mean	*P25*	*P50*	*P75*
Europe	172	72	1.11	0.99	1.10	1.22
USA	122	83	1.55	1.32	1.54	1.72

Notes: Mean, P25, P50, P75: The mean, 25th, 50th (i.e. the median) and 75th percentile of thedistribution. *% Papers from univs*: A rough estimate of the percentage of the total university article output from a country/region published by the universities in the set of 386 world universities. Both percentages are rough estimates, as the number of articles published by the *total* collection of universities in Europe or the USA is not exactly known in this study.

Table 4. Analysis disciplines in which universities are 'world leaders'

Indicator	*All universities*		*Very best 25 % universities*	
	Europe	*USA*	*Europe*	*USA*
Number of universities	172	122	43	31
Among the world top 10 % universities in a discipline				
❖ No (%) universities with at least one 'top' discipline	44 (26 %)	99 (81 %)	29 (67 %)	31 (100 %)
❖ Average number of 'top' disciplines per university	1.8	5.1	2.1	9.3
Among the world top 25 % universities in a discipline				
❖ No (%) universities with at least one 'top' discipline	112 (65 %)	119 (98 %)	43 (100 %)	31 (100%)
❖ Average number of 'top' disciplines per university	3.2	8.4	5.4	12.3

5.4 A country's degree of concentration within the academic research system versus its overall research performance

One of the recommendations made by Lambert and Butler (2006) is to establish more concentration of funding and research activities in a limited number of 'top' universities. The underlying assumption is that more concentration leads to a better performance of the system as a whole, and the high performance of US institutions is a case in point. In order to further analyse the statistical relationship between a country's degree of concentration of research among its universities and its overall research performance, Pratt's Concentration Index of published articles among its universities was calculated for each country. This measure ranges between 0 (no concentration, i.e. all universities publish the same

number of papers) and 1 (total concentration, all papers are published by one single university).

Figure 2 plots for major European countries a country's Pratt's Concentration Index (on the horizontal axis) against the normalized citation impact of the papers published by its universities with at least 500 published articles during 1997–2004 (on the vertical axis). US universities are not included in this analysis, since the study collected for the USA data on a limited number of top institutions only. Countries showing a relatively low Pratt's Index are the Netherlands, Switzerland, Portugal and Hungary. The UK, Austria and the Czech Republic show the highest value of this concentration index. This figure shows that there is apparently no linear correlation between these two variables. The Pearson correlation coefficient amounts to 0.06 (not significant at $p = 0.01$).

These findings illustrate that the relationship between a country's degree of concentration of academic research activities among its universities and its overall performance is complex. In Europe there is no clear tendency that national academic systems showing more concentration of research activities among its universities, generate — as a whole — a higher citation impact per paper than national systems in which the article output is more evenly distributed among academic institutions. Although this issue needs to be analysed in more detail, this outcome itself may be of interest in the debate about the effectiveness of national research policies aimed at establishing greater concentration of research activities among universities in a national academic system.

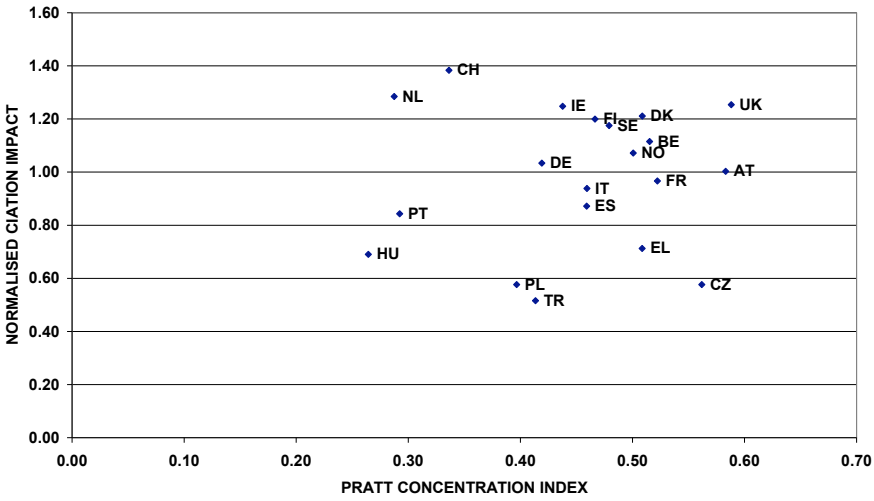

Figure 2. Pratt Index versus normalised citation impact for major European countries (universities with > 500 papers during 1997-2004)

Note: AT: Austria; BE: Belgium; CZ: Czech Rep; CH: Switzerland; DE: Germany; EL: Greece; ES: Spain; FI: Finland; FR: France; HU: Hungary; IE: Ireland; IT: Italy; NL: Netherlands; NO: Norway; PL: Poland; PT: Portugal; SE: Sweden; TR: Turkey; UK: United Kingdom.

5.5 Rankings per research field versus rankings for all fields combined

Rankings of world universities are normally based on indicators calculated for an institution as a whole, combining all research fields in which it is active. In order to illustrate how a ranking of universities based on their bibliometric scores in a particular re-search field may differ from that based on indicators for a university as a whole, this section presents an analysis of one important medical subfield: oncology.

The field 'oncology' was delimited in the following way. In a first step all papers were selected that were published in journals that were included in the WoS journal category oncology. This category contains specialist journals in the subfield. But oncologists also publish papers in more general journals or in specialist journals covering other subfields. Therefore, in a second step the set of papers in specialist journals was expanded. Oncology-related papers were added that were published in journals not included in the ISI journal category oncology but, for instance, in multidisciplinary journals such as Science, Nature, in more general medical journals such as the Lancet and New England Journal of

Medicine, and in journals covering other specialties. These are denoted below as additional oncology papers.

Oncology-relatedness was measured through citation relationships in the following way. From the total WoS database all papers were selected satisfying the following two criteria:

1. At least 10% of documents cited in a paper were published in one of the specialist journals in the WoS journal category oncology.
2. These documents were published in journals of which at least 2% of papers satisfied criterion 1.

Merging the papers in the WoS journal category oncology and the additional papers into one set, the percentage of papers in journals included in the WoS journal category oncology accounted for about 42% of the number in the total set. This percentage was stable over the years. The number of papers in the total set increased from about 39,000 in 1997 to 49,000 in the year 2004. For further details the reader is referred to López-Illescas et al (2007).

Table 5 presents a ranking of the top 25 universities based on the total number of papers they published — in all disciplines — and a ranking according to the number of papers published in the subfield oncology. Data relate to the time period 1997–2004, and to the set of 386 universities publishing at least 5,000 papers during this time period. Table 5 shows that several universities move a significant number of positions up- or downwards in one ranking compared to the other.

Table 5. Rankings based on number of papers in all fields combined and in Oncology

	All Fields Combined			*Oncology*		
Rank	***University***	***Nr Publ / Year in all fields combined***	***Rank Oncoloy***	***University***	***Nr Publ / Year in Oncology***	***Rank in all fields combined***
1	Harvard	9,594	**2**	Univ. Texas - Houston	1,009	68
2	Tokyo	6,445	**8**	Harvard	981	1
3	Toronto	5,104	**4**	Johns Hopkins	483	9
4	Univ. Calif - Los Angeles	4,993	**10**	Toronto	459	3
5	Univ. Washington - Seattle	4,922	**9**	Karolinska Inst Stockholm	451	52
6	Univ. Michigan - Ann Arbor	4,720	**14**	Univ. Calif - San Francisco	415	25
7	Kyoto	4,550	**27**	Penn	389	13
8	Cambridge	4,544	**120**	Tokyo	387	2
9	Johns Hopkins	4,483	**3**	Univ. Washington - Seattle	366	5
10	Univ. Coll London	4,301	**21**	Univ. Calif - Los Angeles	361	4
11	Stanford	4,281	**25**	Pittsburgh	328	29
12	Oxford	4,223	**76**	Wien	327	44
13	Univ Penn	4,134	**7**	Erasmus Univ. Rotterdam	322	135
14	Osaka	3,991	**24**	Univ. Michigan - Ann Arbor	312	6

Table 5. (Continues from previous page) Rankings based on number of papers in all fields combined and in Oncology

	All Fields	*Combined*			*Oncology*	
Rank	***University***	***Nr Publ / Year in all fields combined***	***Rank Oncoloy***	***University***	***Nr Publ / Year in Oncology***	***Rank in all fields combined***
15	Univ. Minnesota – Minneapolis-St Louis	3,987	**26**	Milano	312	37
16	Univ. Wisconsin - Madison	3,930	**49**	Duke	285	31
17	Cornell	3,845	**31**	Ruprecht Karls Univ Heidelberg	281	65
18	Columbia	3,789	**23**	Univ. S. Calif	280	64
19	Univ Calif-Berkeley	3,722	**243**	Baylor Coll. Med	264	109
20	Univ. Calif San Diego	3,697	**58**	Maximilians Univ Munchen	259	34
21	Tohoku	3,681	**72**	Univ. Coll. London	257	10
22	Imperial Coll. London	3,377	**75**	Univ. N Carolina - Chapell Hill	257	39
23	Yale	3,365	**43**	Columbia	253	18
24	Florida	3,363	**111**	Osaka	249	14
25	Univ Calif – San Francisco	3,320	**6**	Stanford	247	11

Table 6 gives for the total set of 386 top universities the Pearson correlation coefficient between the number of articles a university published in all fields combined on the one hand, and the number of published papers in oncology on the other. In addition, it gives the mean, 25th, 50th (median) and 75th percentile of the distribution of the absolute number of positions a university moved in one ranking compared to the other (abs. rank diff.) across universities. Table 6 shows that the mean number of positions universities move in one ranking compared to the other amounts to 103; 25% of universities move at most 28 positions, half of universities move at least 76 positions, while another 25% move at least 140 positions.

Table 5 shows that the position of US universities is less dominant in the oncology ranking than it is in the ranking based on publication counts in all fields combined. This is consistent with the finding presented above that European universities do carry out top research in at least some disciplines, but that the number of disciplines in which they are among the top in the world is lower that of US academic institutions. In other words, the top of US universities is broader, and this leads to higher values of bibliometric indicators — especially publication counts — if these are calculated for a university as a whole. The empirical data presented in this section relate to one field only. But if the interpretation of the outcomes is valid, one would expect that they represent a general pattern, and that generally in analyses of research fields or disciplines the position of US universities tends to be less dominant than it is in an overall ranking according to total publication counts in all fields combined.

Table 6. Comparison Rankings based on number of papers in all fields combined and in Oncology

	Pearson R	***Abs Rank Diff***			
Universities		Mean	P25	P50	P75
Top 386	0.71	103	28	76	140

Note: *abs rank diff*: absolute number of positions a university moved in one ranking compared to the other. P25, P50, P75: the 25th, 50th (=median) and 75th percentile of the distribution of the variable *abs rank diff* among universities.

5.6 General versus specialised universities

It is useful to distinguish between general and specialised universities. General universities cover a wide range of scientific-scholarly disciplines. A typical example is a university that offers courses and carries out research in all domains of science and scholarship. Specialised universities are mainly active in a limited number of disciplines. Often — but definitely not in all cases — their name reveals the disciplines on which they focus. Typical examples are technical, medical, and agricultural universities.

Although general universities show less concentration of research activities among disciplines than specialised universities, they do not necessarily have the same level of activity in all disciplines. They may be more active in some disciplines than in others, and their research profile may reveal a certain specialisation, though not as pronounced as in specialised universities. In practice, it is very difficult to draw a sharp borderline between general universities with a certain specialisation on the one hand, and specialised universities on the other. The transition from the first to the second group is fluent.

This section analyses disciplinary specialisation within a university, i.e. the extent to which its research papers are evenly distributed among research disciplines, or whether there are particular disciplines on which a university focuses its research activities. Figure 3 plots, for each of the 386 universities with at least 5,000 papers during 1997–2004, their disciplinary specialisation index measuring the degree of concentration of their published articles among disciplines (on the horizontal axis), against the normalised citation impact of its papers (on the vertical axis). In order to obtain an impression of differences across countries, universities from the Netherlands, Germany, Sweden, UK and USA are indicated by special symbols.

Figure 3 reveals that in the total set of 386 universities there is no simple relationship between these two variables. The line drawn in this figure is the linear regression line. The Pearson and Spearman rank correlation coefficients are –0.06 and –0.10, respectively, and are not significant at $p = 0.01$. In this set of 386 world universities, general universities showing a rather even distribution of research papers among disciplines, and specialised universities having their article output concentrated in a limited number of disciplines (regardless of which ones), show statistically similar citation impacts. It must be noted that smaller specialised universities such as the London School of Economics are not

included in the analysis, since the number of papers published by this institution does not exceed the threshold of 625 papers per year.

The impact measure plotted in Figure 3 relates to a university's total article output in all fields combined. More information can be obtained from an analysis by discipline, addressing the question: Do specialised universities in their fields of specialisation perform better than general universities do in the same fields? Specialisation is defined here from a disciplinary perspective, in terms of the distribution of a university's research articles among 15 disciplines, listed in Table 2.

For each university, the normalised citation impact was calculated for all papers in each of the 15 disciplines separately. In order to correct for large differences in universities' normalised citation impact across countries, a further normalisation of the citation impact indicator was carried out, by calculating per discipline the ratio of the citation impact of a university from a particular country and the mean citation impact across all universities in that country. This 'double'-normalised impact indicator was correlated with the publication activity index, expressing the institution's specialisation in a discipline, based upon the distribution of its papers among disciplines compared to the world distribution. Only universities with at least 50 papers in a discipline were included in the correlation analysis for that discipline.

Table 6 presents the outcomes of this analysis. In four disciplines a significant correlation was found between citation impact per paper and degree of specialisation (publication activity index): in biological sciences primarily related to humans, clinical medicine, molecular biology and in physics, with Pearson coefficients of 0.24, 0.23, 0.41 and 0.17, respectively. In all other disciplines the correlations were not significant at $p = 0.01$.

These outcomes await further interpretation. The disciplines in which a significant, positive correlation was found embrace typical 'big science' fields, and perhaps the outcomes show that the concept of 'critical mass' in research activity is more relevant in 'big science' than it is in other domains of science and scholarship. It needs emphasising, however, that this analysis focuses on specialisation across rather broadly defined disciplines, and that it does not take into account specialisation within a discipline. A more detailed study could further analyse differences across countries and subject of specialisation.

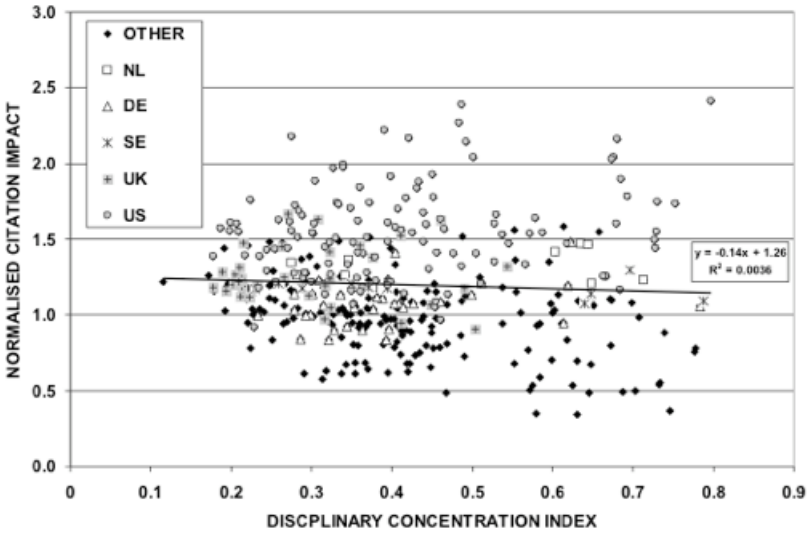

Figure 3. 386 World Universities' disciplinary specialisation index versus their normalized citation impact

5.7 Collaboration networks of universities using social network analysis

In order to analyse the structure of a national academic system and highlight the position of individual universities, maps based on network analysis are particularly useful. Such maps allow one to identify the best research universities in their national or regional environment, based on a series of bibliometric and network indicators (Calero-Medina and Moed, 2006). Institutions are not ranked on the basis of one single indicator. Instead, a social network analysis is applied to represent relations between universities based on co-authorship, and to identify patterns of co-publication activity. The novelty of this approach is that one may identify not only the best research universities based on a series of bibliometric indicators, but also analyze the way in which universities collaborate, and their position in a global collaboration map.

The institutions were characterized by the following properties:

- The country or geographical region in which the university is located.

- The number of papers published by the university. The size of the circles or triangles representing a university indicates the number of papers it published during the time period considered.

- The field normalized citation impact indicator of a university's publication output. Triangles represent universities that are among the top 25% in terms of the normalized citation impact of the articles it published.

The thickness of a connecting line indicates the strength of the co-publication relationship among a pair of universities. This strength is expressed by Salton's Index, defined as the number of co-publications between two universities, divided by the square root of the product of the number of papers published by each university.

Figure 4 shows the global collaboration network among the top 100 world universities in terms of the number of articles published during 1997–2004. It displays only co-publication links of which the strength exceeds 0.02. Applying this threshold, about 12% of the co-publication matrix was taken into account. The map reveals the bridge function of Canadian universities between Europe and the USA, and the central position of the University of Tokyo that shows strong collaboration links with European and North American universities. The collaboration patterns among European universities clearly reflect the importance of national collaboration. This may suggest that the European Research Area is not (yet) as strongly integrated as the research activities carried out in the USA.

In a reasonably easy but still reliable way one generates in just one network map an overall picture of the set of universities (national, regional or worldwide level) based on bibliometric performance measurements. This can be carried out both at an aggregate level of a university as a whole, but also per research field, in order to identify field specific characteristics.

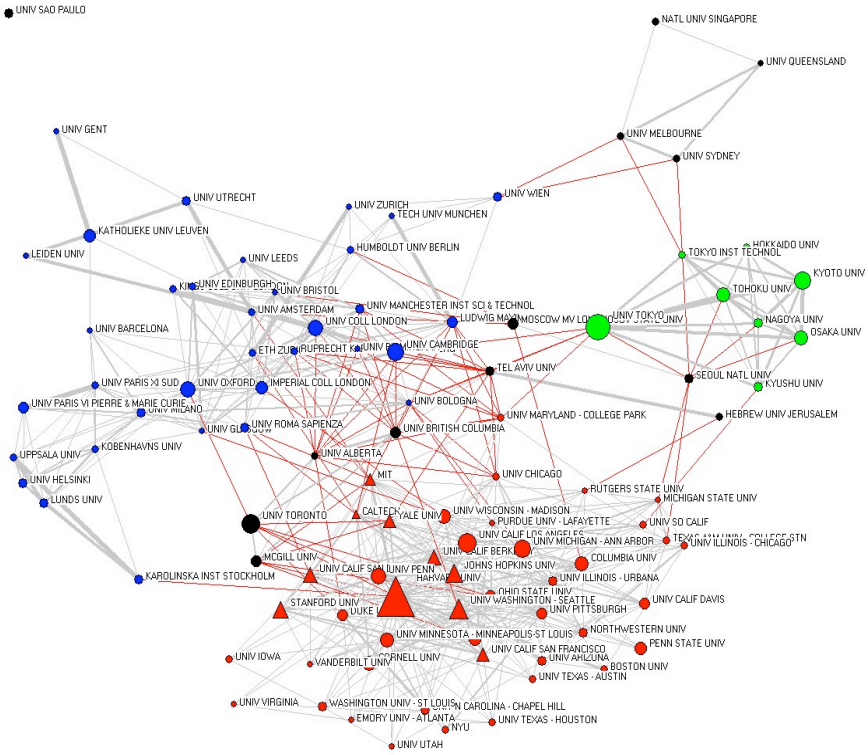

Figure 4. Collaboration network of top 100 world universities

Note: The figure shows the global collaboration network among the top 100 world universities in terms of the number of articles published during 1997-2004. It displays only co- publication links of which the strength (Salton's Index) exceeds 0.02. The thickness of a connecting line indicates the strength of the co-publication relationship among a pair of universities. Each country or region has its own colour in the map. The size of the circles or triangles representing a university indicates the number of papers it published during the time period considered. Triangles represent universities that are among the top 25 per cent in terms of the normalised citation impact of the articles it published.

5.8 Concluding remarks

More detailed empirical studies should be made of the structure of European, US and other (supra-) national academic systems. A first research topic is the extent to which these systems are structured according to a concentration or a distributed model. Therefore, one should also analyse the performance of other US universities than the 122 studied in this paper, and of academic institutions in other countries. A key question is which of the two is the most appropriate in the various countries, especially which model provides the most optimal conditions

for ‘top’ research. This complex, policy relevant question awaits further study.

A second way to further examine the structure of national academic systems is to produce for a number of countries maps of the type presented in Section 5.7 based on social network analysis. Such maps would reveal the positions of individual universities and their relationships within a national academic system. A challenge would be to compare the structures that are obtained for the various countries with one another, to develop a classification system of these network structures, and to characterise countries accordingly. In addition, it would be useful to further characterise the role of individual universities in terms of whether they have an international, national or local orientation.

A practical implication of the findings presented in Section 5.6 is that it would be appropriate to compile and publish rankings of universities per research field or discipline. In addition, one should consider in rankings based on indicators calculated for all fields combined to add for each university the value of its disciplinary specialisation index, and to indicate for more specialised universities the discipline(s) in which they specialise. This would substantially enhance the information content and utility of the rankings.

The publication data for the universities analysed in this paper were *not* verified by representatives of the institutions, except in a few cases. A main future task will be to find ways to enable them to verify the data. The bibliometric data used in this study focus on the 'output' side of research. It should be combined with other publicly available, verified or certified information, reflecting aspects of the 'input' side, including per discipline at least the number of students and various categories of research staff, and the amount of public funding. Although these ‘input’ measures partly reflect ‘output’ categories such as research quality as well – for instance, ‘good’ institutions tend to attract more funding than less good ones – their use in statistical analyses is indispensable, and will enrich the comparative analysis of national academic systems.

It is essential that these data are not only available at the level of a university as a whole, but at least also by discipline, in order to relate ‘output’ to ‘input’ at the level of disciplines. Therefore, the mismatch that currently exists between disciplinary categorisations at the output and the input side needs to be solved (Luwel, 2004). In this way, a public information system on world research universities can be built, that is not only useful for the general public, but also constitutes a database for further research on research performance and its determinants.

References

Bookstein, A., and Yitzahki, M. (1999). Own language preference: A new measure of "relative language self-citation". *Scientometrics*, 46, 337–348.

Calero-Medina, C. and Moed, H.F. (2006). Depicting the landscape of research universities. Paper presented at the Ninth International Conference on Science and Technology Indicators, Leuven (Belgium), 7-9 September 2006.

Diamond, N., and Graham, H.D. (2000). How should we rate research universities? http://www.physics.northwestern.edu/graduate/Graham_Diamond.html.

Egghe, L. and Rousseau, R. (1990). Introduction to Informetrics. Amsterdam: Elsevier.

Garfield, E. (1964). The Citation Index – A new dimension in indexing. *Science*, 144, 649–654.

Garfield, E. (1979). Citation Indexing. Its theory and application in science, technology and humanities. New York: Wiley.

Goldberger, M.L, Maher, B.A, and Flatteau, P.E. (1995). Research-doctorate programs in the United States: Continuity and change. National Academy Press.

Lambert, L. and Butler, N. (2006). The Future of European Universities: Renaissance or Decay? London (UK): Centre for European Reform (CER), ISBN 1 901 229 67 X.

López-Illescas, C., Moya-Anegón, F., and Moed, H.F. The actual citation impact of European oncological research. Manuscript submitted to *European Journal of Cancer*.

Luwel, M. (2004). The use of input data in the performance analysis of R&D systems. In: Moed, H.F., Glänzel, W., and Schmoch, U. (2004) (eds.). *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*. Dordrecht (the Netherlands): Kluwer Academic Publishers, 315–338.

Matia, K., Amaral, L.A.N., Luwel, M., Moed, H.F., Stanley, H.E. (2005). Scaling phenomena in the growth dynamics of scientific output. *Journal of the American Society for Information Science and Technology* ***56*****,** 893-902.

Moed, H.F., de Bruin, R.E., and van Leeuwen, Th.N. (1995). New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. *Scientometrics*, 33, 381–442.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht (Netherlands): Springer. ISBN 1-4020-3713-9, 346 pp.

Moed, H.F. (2006). Bibliometric Rankings of World Universities. CWTS Report 2006-01. Leiden (Netherlands): Centre for Science and Technology Studies. Available at: http://www.cwts.nl/hm/bibl_rnk_wrld_univ_full.pdf.

Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. Washington D.C.: National Science Foundation.

SJTU (2007). Academic Ranking of World Universities -2007. Shanghai Jiao Tong University, Institute of Higher Education, published on 15 August 2007, available at http://www.arwu.org/rank/2007/ranking2007.htm.

THES (2007). The Times Higher World University Rankings 2007. The Times Higher Education Supplement, available at http://www.thes.co.uk/worldrankings/.

Van Raan, A.F.J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36, 397–420.

Van Raan, A.F.J. (2005). Challenges in Ranking of Universities, available at www.cwts.nl/cwts/AvR-ShanghaiConf.pdf.

Visser, M.S, Calero-Medina, C and Moed, H.F. (2007). Beyond rankings: The role of large research universities in the global scientific communication system. In: Torres-Salinas, D. and Moed, H.F. (eds.). Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, 25-27 June, 2007, Vol II, 761-765.

6 Combining Mapping and Citation Network Analysis for a better understanding of the scientific development: The Case of the Absorptive Capacity field

Clara Calero-Medina, Ed C.M. Noyons
(published in the Journal of Informetrics, 2: 272-279, 2008)

6.1 Introduction

In terms of citations fields or areas of specialization are not just 'formless' sets of articles. On the contrary, they represent sets of papers with a particular structure that emerges from the citation practices of the researchers active in that field. It emphasizes the importance and visibility of certain theoretical and methodological approaches while marginalizing others. We could say that citation practices represent a "knowledge-construction" process that outlines the manner we think about and engage with our research. The emergence of trajectories (Dosi, 1982) implies that the evolution of knowledge is not random.

In every scientific field there are key concepts that set the base for theoretical developments through the years. As De Nooy, Mrvar & Batagelj (2005) pointed out, citation analysis may focus on the identification of specialties, the evolution of research traditions, and changing paradigms. Researchers from the same specialty tend to cite each other in order to position their work in the field based on previous knowledge. Scientific knowledge is assumed to increment over time following a "smooth path", the papers that introduce important new insights are cited until new results modify or contradict them. The scientific revolutions, sudden paradigmatic changes resulting from new insights (Kuhn 1969), are reflected by abrupt changes in the citation network.

The objective of our study is analyzing the influence of the introduction of a new concept on a research field through the analysis of scientific publications.

1. How the diffusion of the concept has taken place through the research literature building over the original Absorptive Capacity concept?
2. Which papers and theories are considered the main research streams of the field?
3. Which papers are essential?

The novelty of our approach is that to answer to these questions we combine bibliometric mapping and citation network analysis. The bibliometric co-word map provides insight into the contents of the publication while two techniques from the citation network analysis recognized the main papers during fifteen years. This is used for the interpretation of groups of citations that may constitute the backbones of a research tradition or the future of the research.

6.2 Data and Methods

Data

In the research field Organization the concept "Absorptive Capacity" (AC) is considered as one of the most important introduced in the last fifteen years. In their study on international transfer, Kedia and Bhagat (1988) first coined the term "Absorptive Capacity". However, the contribution by Cohen and Levinthal (1990) is generally accepted as the founding paper. It defined AC as "the ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends". Cohen and Levinthal put R&D at the center of firm's innovative processes by linking it to both learning and innovation (Volberda, Foss & Lyles, 2006). Nevertheless, the Cohen and Levinthal contribution did not emerge out of the blue, and the AC theme overlaps with other themes and fields, such as cognition, knowledge and dynamic capabilities[1]. The theoretical development of AC ranges from the psychological emphasis on cognition and learning to the focus of economics on innovation and competition to the sociological orientation towards co evolution. (Volberda, Foss & Lyles, 2006)

This influential publication has received more than 1,500 citations (up to 2007) in papers published in journals processed for the Web of Science (WoS) published by Thomson Scientific, and as Figure 1 shows, the attention is growing. Recently there have been two main efforts for reviewing the absorption of the concept of Absorptive Capacity in the literature of Organizational Theories (Foss, Lyles & Volberda (in press) and Lane, Koka & Pathak (2006)). These papers point out the main streams in the field of Organization related to Absorptive Capacity: organizational learning, innovation, the knowledge-based view of the firm, dynamic capabilities, co-evolution and managerial cognition. Some of these experts were involved in the validation of the results of our study.

[1] In economics, the idea of "learning to learn" introduced by Stiglitz' (1987) is clearly a precursor of AC, as is David's (1975) work of localized technological progress.

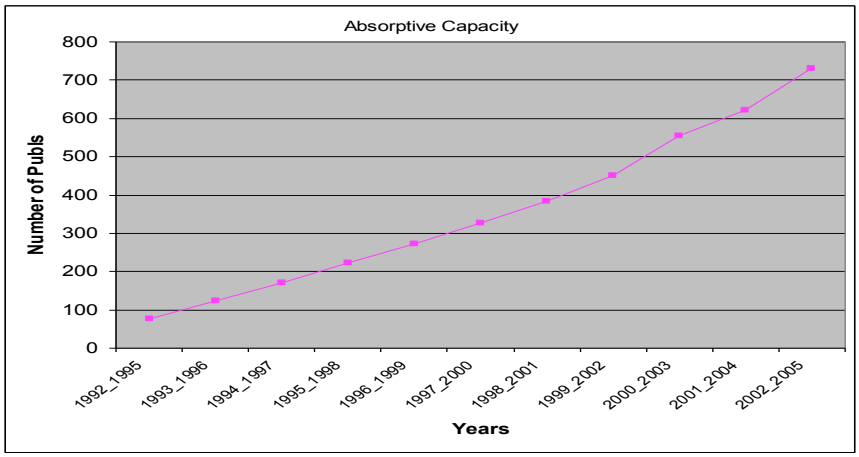

Figure 1. Number of Publications citing Cohen & Levinthal (1990) during the period 1992-2005 (Source: Web of Science)

The data set consists of the 1213 publications citing Cohen and Levinthal (1990) up to 2005. The publications were extracted from journals covered by the Web of Science. We define this set of publications as the 'Absorptive Capacity field'.

Bibliometric Map Analysis

The first step was to get a general overview of the Absorptive Capacity field. We map the structure of all publications citing Cohen and Levinthal (1990) with a bibliometric mapping method based on keyword co-occurrences. With this method we created a 2-dimensional graph with sub-domains representing topic clusters. The topic clusters were created by applying a co-word analysis to the keywords in the *citing* publications (Noyons, 1999). We collected the keywords of these 1213 publications to assess the contents of the field. These keywords were extracted from the bibliographic fields *keywords plus* and *author keywords*. The former are keywords automatically assigned by Thomson to individual publications on the basis of cited reference information. The latter are the keywords assigned to publications by the authors.

Of the 94 most frequent keywords (with 20 or more occurrences) we selected the 83 most relevant and discriminative. This selection was done by experts in the field of *Absorptive Capacity* in close collaboration with the authors. With these 83 keywords, we calculated the number of times they co-occured in publications. With this information, we applied a

(hierarchical agglomerative, complete linkage) cluster analysis and identified 11 clusters of topics (keywords). We refer to these keyword or topic clusters as sub-domains in the field. Using these topics, we were able to assign individual publications to sub-domains. In addition, we defined the overlap between the clusters, with the publications present in more than one sub-domain. This overlap provides input for the cosine similarity measure between sub-domains. Multi Dimensional Scaling (MDS) was applied to the obtained similarity values, and this application yielded a map of sub-domains. The distances between sub-domains represent their cognitive similarity in terms of common publications. The closer they are in the map, the more similar. The validation (and the label) for each sub-domain was provided by the above mentioned field experts. These labels are compiled to represent application areas of AC. As such they are not directly retrieved from the publication data but rather created by the experts and referring to actual research areas. For further details of the mapping methodology, we refer to Noyons (1999).

This part of the analysis gave us a first overview of the field. We could identify the sub-domains that attract more publications and their growth rate in terms of number of publications over the period. But we still didn't know anything about the publications behind these sub-domains.

Publication content labeling

The next step was to label each of the 1213 publications citing Cohen & Levinthal (1990) with the sub-domain(s) to which they belong. Thus we were able to classify the publications with respect to content.

Citation Network Analysis

Subsequently, we created a citation network based on the citation links between the 1213 papers. Citation network analysis began with the study by Garfield, Sher & Torpie (1964) of Asimov's history of DNA. This study showed that there was "a high degree of coincidence between an historian's account of events and the citation relationship between these events". In our study, we carried out a citation network analysis to investigate the processes of the diffusion of the concept of Absorptive Capacity and the theories around it. The citation analysis allowed us to view the structure of part of the Absorptive Capacity literature that had emerged from current citation practices and showed how this emergent structure elevated certain approaches and marginalized others. In this context, following Small (1978), a cited document stands for a concept. Highly cited documents have a significant content that is shared by a community of scientists. A publication often cited may be seen as a

"concept symbol" that represents an author's orientation to a community of scientists or an approach to a topic (Moed, 2005).
The citation network enabled us to study the data from two perspectives in time. In the evolution of knowledge, phases of consolidation of past results coexist with exploration of new approaches. The techniques of longitudinal network analysis, like main path analysis, allowed us to unravel the dynamics of convergence and divergence between streams of investigation (Ramlogan et al., 2007). It shows the change over time of the connectedness of the system. The second perspective is a cross-sectional look at the state of the literature in 2005 through the identification of important publications based on a 'hubs' and 'authorities' analysis (Kleinberg, 1999). These two perspectives were important because they highlighted different parts of the citation network.

Main Path

If knowledge flows through citations, a citation that is needed in paths between many articles is more important than a citation that hardly plays any role for linking articles (De Nooy, Mrvar & Batagelj, 2005). Among all possible "chains" of citations from the most recent records to the oldest, the network algorithm computes the paths that are most frequently encountered, and these can be regarded as the backbones of a research tradition (Hummon & Doreian, 1989, 1990; Hummon & Carley, 1993; Batagelj, 2003; De Nooy, Mrvar & Batagelj, 2005). These results identify the path that is most frequently used to 'walk' from the present to the pass (that is back in time) in a 'field' of papers: the 'main path'. We stress that this method does not involve the absolute count of maximum number of citations received, but the simultaneous computations of all the possible paths through the whole dataset and the choice of the one that is the most frequently encountered through time (Mina et al., 2007).
As Batagelj showed in 2003 with the example in SOM (self-organizing mapping) literature, a "main subnetwork" can be extracted applying a similar procedure as the main path analysis. The main subnetwork contains not only the main path but also other important branches from the citation network provide rich information about the development of a field. In this paper though, we were more interested in showing a new methodology that combines different approaches. In order to keep things as simple as possible for a better understanding of the methodology, we will apply only a main path analysis. We are quite aware though of the additional information that the main subnetwork can provide us with.

As an illustration of how the main path is extracted from a citation network we have prepared a simple example. Figure 2 shows a citation network of eleven fictitious papers ordered in time from top to bottom.

The vertices (circles) represent papers and the arcs (arrows) indicate cited by. A Source Vertex is an article that is not citing within the data set (*P1-2003*). A Sink Vertex is an article that is not cited within the data set (*P11-2007*). In the network terminology, a path is a walk in which no vertex or arcs in between the source and the sink vertex occurs more than once. For extracting the main path from the citation network we first computed the 'traversal weights'. The traversal weight measures the number of times that a link between articles was involved in connecting other articles in a citation network. The thickness of the arcs in Figure 2 shows the traversal weight measure. In a citation network, a main path network following the Search Path Count (Batagelj, 2003) is constructed starting from the source vertex and selecting at each step in the end vertex the lines with the highest weight, until the sink vertex is reached. Starting from the source paper (*P1-2003*), the main path algorithm chooses the next link in the path as the outgoing link with the highest traversal weight *(P2-2004),* from this one the highest link drives us to *P5-2005*, from here to *P9-2006* and *P10-2006*, to finish in the sink vertex *P11-2007*. By repeatedly applying this choice rule, we defined a path through the network that follows a structurally determined most used path.

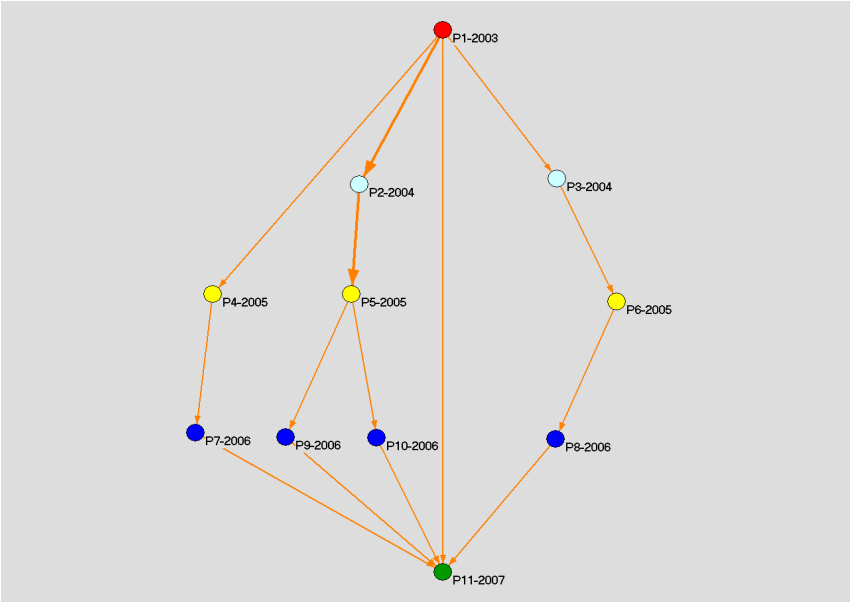

Figure 2. Traversal Weights in a citation network.

The main path, chosen on the basis of the most used path identified the main stream of the Absorptive Capacity literature between 1990 and 2005, having the Cohen and Levinthal (1990) as source paper. The main path analysis was conducted with the software package Pajek.

Hubs & Authorities

Research in bibliometrics and in context of hypertext and the www are concerned with the identification of important nodes in networks. The famous Garfield's impact factor (Garfield, 1972) is basically a ranking measured based on a pure counting of the in-degrees nodes in a journal citation network. Not happy with this measure Pinski and Narin (1976) and Geller (1978) developed an algorithm that considered not only the number of citations from one journal to the other but also the prestige of the citing journal. Journals that receive many citations from prestigious journals are considered highly prestigious themselves. By iteratively passing prestige from one journal to the other, a stable solution is reached which reflects the relative prestige of journals (Bollen, 2006). This way of measuring prestige is behind the PageRank algorithms to evaluate the status of web pages. First developed by the founders of the Google search Engine Brin and Page ((Brin & Page, 1998) and (Page at al., 1998)). The PageRank is calculated by an iterative algorithm which propagates prestige values from one web page to another and converges to a solution (Pillai et al., 2005)

In the same period that Brin and Page, Kleinberg (1999) was also working on an algorithm to increase the effectiveness of Web search engines using the concepts of *hubs and authorities*. An authoritative publication, in our case, is one that many other publicatons cite to. But, this idea can be reinforced by observing that citations from all publications aren't equally valuable – some publications are better *hubs* for a given publications. Hubs & Authorities are formal notions of structural prominence of vertices in directed graphs (Brandes & Willhalm, 2002). Kleinberg developed an iterative algorithm for computing hubs and authorities. Hubs and authorities stand in a mutually reinforcing relationship: a good authority is a publication that is cited by many good hub, and a good hub is a document citing to many good authorities. He showed examples where the algorithm could help to filter out irrelevant or poor quality documents (they would have low authority scores) and to identify high-quality documents (they would have high authority scores).

From our perspective making the classification in hubs and authorities is a very useful tool to understand the role playing by the publications in this citation network. From the hubs/authorities perspective for a

publication being both a hub and an authority at the same time is "the best" position: having a lot of influence (authority) but also being influenced by the best (hub). We could say that in terms of knowledge flow and the quality of knowledge used is a good position. This is the reason why on this study we decided to use Kleinberg's algorithm for identifying the main publications in this citation network. Batagelj adapted for the software Pajek Kleinberg's hubs/authorities algorithm (Batagelj & Mrvar, 2006).

For each paper (p) in our citation network we computed two weights: hub weight (h_p) and authority weight (a_p). They show the strength of a given paper as an authority and/or a hub. Weights are computed according to the citation network (M) by solving the eigenvector problem of matrices MM^T (hubs) and M^TM (authorities), where M is the citation matrix (Kleinberg 1999). Paper x is a better hub than paper y if $h_x > h_y$. Paper x is a better authority than paper y if $a_x > a_y$. The hubs and authorities analysis was conducted with the software package Pajek.

6.3 Results

The Absorptive Capacity Bibliometric Map

Figure 3 shows the map of the Absorptive Capacity Field. As explained above, it is the result of clustering keywords (sub-domains) and mapping these sub-domains in a two-dimensional figure, with the size of each sub-domain indicating the number of publications represented and the colour (grey scale) of each sub-domain indicating the growth in the number of publications until 2005 (black: fast growth; grey: growth around average; white: growth below average). The growth rate is calculated by the development of the share of a sub-domain within the entire field. For two 7-years periods we compared these shares. If the share increased in the most recent period with more than 20% it was indicated as a significant growth.Sub-domains closer to one another have more publications in common than sub-domains that were further apart.

Most of the studies in Absorptive Capacity are focused on R&D rates in various industries (sub-domain 9), inter-organizational and managerial antecedents (sub-domain 1 and 2). Fast growing areas of Absorptive Capacity (the black circles) appear to be studies on ***Knowledge flows and capabilities*** (sub-domain 6), the impact of Absorptive Capacity on ***Technological innovation and firm performance*** (sub-domain 10), and the effects of relational (trust) versus formal ***Governance modes*** (sub-domain 7) on Absorptive Capacity. Figure 3 also shows that

Organizational Innovation (sub-domain 11) and ***Realized Absorptive Capacity*** (sub-domain 3) is underrepresented according to the experts.

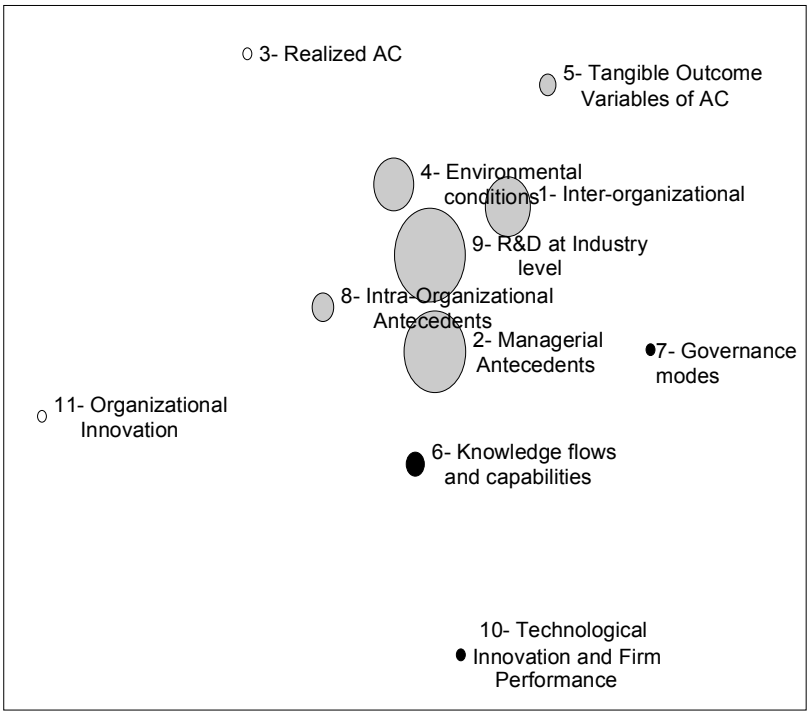

Figure 3. Bibliometric map of the field of Absorptive Capacity.

Citation Network Analysis

Main Path

The main path (Figure 4) shows the main track followed by the researchers in this field to explain industrial innovative processes. The nodes (circles) of the graph represent the publications, the presentation is ordered in time from top to bottom (from 1990 until 2005), the colors (grey scale) represent the year of the publication, and the thickness of the lines relate to the traversal weights. The publications arre labeled with the first author's name, publication year and between parentheses appear the sub-domain/s' number.

Figure 4 illustrates that there were just a few publications that constitute the main stream on the Absorptive Capacity literature. The analysis of the papers along the backbone in Figure 4 provides the sequence of the papers. As we can see in the graph, these papers are strongly focused on the main sub-domains of the map (1, 2, 4, 9), as we can expect from a map with such a central and big sub-domains. However, in 2001 the paper from Ahuja and Katila included some notions relating with one of the small sub-domain "***Organizational Innovation***". In 2002 the paper from Zahra and George identified key dimensions of absorptive capacity and offered a reconceptualization. This analysis was based on sub-domains 1 and 2, but also on 5 and 6. The nodes on the bottom of the diagram were a sample of the state of the art at 2005.

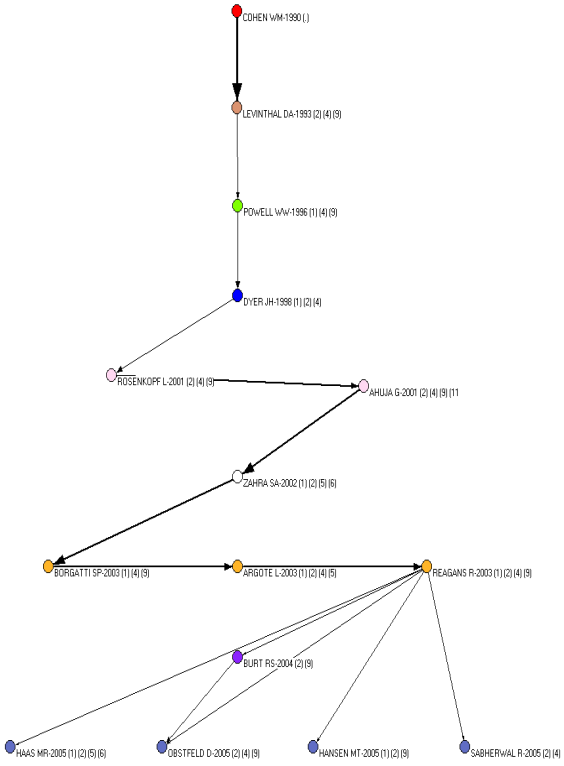

Figure 4. Main Path component of the Absorptive Capacity Field
(vertical dimension represents the publications year, horizontal dimension locates the publications in the same year)

Hubs and Authorities

Table 1 shows the 20 for the hubs and authorities analysis (as explained above). These papers are considered the main hubs and authorities from the citation network. These two lists show the papers that cited the most in general and, in particular, were most cited in our network. As it explained also above, a good hub is a paper that points to many good authorities, and a good authority is a paper that is pointed to by many other good hubs. As we can observe from the date of the publications, the authorities are older paper than the hubs.

The hub papers are in many cases broad literature reviews, but in a few cases these papers attract a lot of attention (i.e., are cited frequently) and thus become authority documents (in bold in Table 1). In this case only one of the three hub/authority papers played a critical role in the main development of the field. The paper from Zahra and George (2002) 'Absorptive capacity: A review, reconceptualization, and extension', that is both an authority as well as a hub, forms part of the basic structure of the main path. The other two papers: one from Kale, P., Singh, H., Perlmutter, H. (2000) 'Learning and protection of proprietary assets in strategic alliances: building relational capital' and the other from Larsson, R., Bengtsson, L., Henriksson, K., Sparks, J. (1998) 'The interorganizational learning dilemma: collective knowledge development in strategic alliances', does not appear in the main path, but further study located this paper in the main subnetwork (as we mentioned previously, in this study we focus only on the main path) as part of the other important paths of the field.

Table 1. Authorities & Hubs

Ranking	h_p	HUB ID	a_p	AUTHORITY ID
1	0.10326	**KALE P-2000 (1) (2) (7) (9)**	0.94172	COHEN WM-1990 (.)
2	0.06417	MARTIN X-2003 (1) (2) (4) (5)	0.12772	SZULANSKI G-1996 (1) (2) (4) (9)
3	0.06124	**LARSSON R-1998 (1) (2) (7)**	0.10879	GRANT RM-1996 (9)
4	0.04842	INKPEN AC-2000 (1) (2) (7)	0.09865	POWELL WW-1996 (1) (4) (9)
5	0.04788	REID D-2001 (2) (7)	0.09707	LANE PJ-1998 (1) (2) (4) (9)
6	0.04775	IRELAND RD-2002 (1) (2) (7)	0.08833	LEVINTHAL DA-1993 (2) (4) (9)
7	0.04703	NIELSEN BB-2005 (1) (2) (4) (6)	0.08572	MOWERY DC-1996 (1) (4) (5) (9)
8	0.04442	MALHOTRA A-2005 (1) (2) (4) (6)	0.07299	DYER JH-1998 (1) (2) (4)
9	0.04374	**ZAHRA SA-2002 (1) (2) (5) (6)**	0.06332	VONHIPPEL E-1994 (4) (9)
10	0.04357	ANDERSSON U-2002 (1) (2) (8) (9)	0.05746	NAHAPIET J-1998 (1) (2) (8) (9)
11	0.04289	SIMONIN BL-2004 (1) (2) (9)	0.05593	CONNER KR-1996 (2)
12	0.04250	CUMMINGS JL-2003 (1) (2) (5) (9)	0.04975	**KALE P-2000 (1) (2) (7) (9)**
13	0.04222	MOLINA LM-2004 (1) (2) (10)	0.04833	**ZAHRA SA-2002 (1) (2) (5) (6)**
14	0.04160	MATUSIK SF-2005 (1) (2) (9)	0.03839	**LARSSON R-1998 (1) (2) (7)**
15	0.04140	ARANDA DA-2002 (1) (2) (6) (9)	0.03769	SIMONIN BL-1999 (1) (2) (4) (6)
16	0.04074	SIMONIN BL-1999 (1) (2) (4) (6)	0.03714	KHANNA T-1998 (1) (2) (8) (9)
17	0.04062	BARRINGER BR-2000 (1) (2)	0.02969	ANDERSON P-1990 (4)
18	0.04017	HOLMQVIST M-2003 (1) (2) (4) (9)	0.02900	GUPTA AK-2000 (1) (2) (4) (9)
19	0.03982	ALMEIDA P-2004 (1) (2) (5) (9)	0.02729	BOWMAN EH-1993 (1) (2) (4) (8)
20	0.03963	JOSHI AW-2003 (1) (2)	0.02689	LIEBESKIND JP-1996 (4) (8) (9)

(bold: papers that are hubs and authorities at the same time)

6.4 Concluding remarks and follow-up research

We think that our results show to information scientists the potential of this new methodology as a tool for unraveling the patterns behind a set of publications representing a field. The combination of bibliometric mapping with citation network techniques enables us to follow the influence of the introduction of a new concept in a specific research field. The use of bibliometric mapping with network analysis as a useful tool in the identification of research groups has also been demonstrated in a previous study of the authors (Calero et al., 2006).

In this study, as an example, we followed the 'intellectual track' of a specific concept, *Absorptive Capacity* (AC), with a high rate of diffusion through the fifteen years of analysis. The bibliometric map identifies the other concepts (in terms field-specific keywords) and the theories associated with the main concept (AC) while two techniques from the citation network analysis recognized the main papers during these years, the articles that influenced the research for some time and linked them

into a research tradition that is the backbone of the 'Absorptive Capacity Field'.

It is important to mention that because of our focus on a specific term, the analysis is based on a few central sub-domains. Of course, the Cohen and Levinthal contribution did not emerge out of the blue, and the Absorptive Capacity topic overlaps with other research themes and fields, such us cognition, knowledge flow and dynamic capabilities as major parts of the field 'Organization'.

Our next goal will be to map and detect all main research streams in a physics-chemistry related field, and particularly to identify the papers considered as authorities and hubs. Bibliometric maps that are not focused on just a specific topic of a field will show many different parts of that field of which the dynamics will be represented by the main path analysis.

Acknowledgments

We thank Prof. dr. Henk W. Volberda, from the Rotterdam School of Management, for evaluating the results and for his helpful comments.

References

Ahuja, G. & Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: a longitudinal study, Strategic Management Journal 22, 197–220.

Asimov, I. (1963). The Genetic Code. New York: New American Library.

Batagelj, V. (2003). Efficient Algorithms for Citation Network Analysis. Preprint Series. Univ. Ljubljana, Inst. Math., 41 (897),1-29.

Batagelj, V. & Mrvar, A. (2006). Pajek: Program Package for Large Network Analysis, University of Ljubljana, Slovenia. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Brandes, U. & Willhalm, T. (2002). Visualization of bibliographic networks with a reshaped landscape metaphor. Joint Eurographics-IEEE TCVG Symposium on Visualization, D. Ebert, P. Brunet, I. Navazo (Editors). http://algo.fmi.uni-passau.de/~brandes/publications/bw-vbnr1-02.pdf.

Bollen, J., Rodriguez, M.A. & van de Sompel, H. (2006). Journal Status. Scientometrics, 69 (3), 669-687.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30, 107-117.

Calero, C., R. Buter, C. Cabello Valdés, E. Noyons (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. Scientometrics, 66 (2),365–376.

Cohen, W.M. & Levinthal, D.A. (1990). Absorptive Capacity – A New Perspective on Learning and Innovation. Administrative Science Quaterly, 35 (1), 128-152.

David, P.A. (1975). Technical choice innovation and economic growth, Cambridge, Cambridge University Press.

De Nooy, W., Mrvar, A. & Batagelj, V. (2005). Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press.

Dosi, G. (1982). Technological Paradigms and technological Trajectories: A Suggested Interpretation of the Determinants and Directions of Technical Change. Research Policy, 11 (3), 147-162.

Foss, N.J., Lyles, M.A. & Volberda H.W. (in press). Absorbing the Concept of Absorptive Capacity: How to Realize its Potential in the Organization Field. Organization Science.

Garfield, E., Sher, I.H. & Torpie, R.J. (1964). The Use of Citation Data in Writing the History of Science. Philadelphia: Institute for Scientific Information.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178, 471-479.

Geller, N.L. (1978). On the Citation Influence Methodology of Pinski and Narin. Information Processing & Management, 14, 93-95.

Hummon, N. & Carley, K. (1993). Social networks as normal science. Social Networks, 15, 71–106.

Hummon, N. & Doreian, P. (1989). Connectivity in a citation network: the development of DNA theory. Social Networks, 11, 39–63.

Hummon, N. & Doreian P. (1990). Computational methods for social network analysis. Social Networks, 12, 273–88.

Kale, P., Singh, H. & Perlmutter, H. (2000). Learning and protection of proprietary assets in strategic alliances: building relational capital, Strategic Management Journal, 21, 217-37

Kedia, B.L. & Bhagat, R.S. (1988). Cultural constraints on transfer of technology across nations: Implications for research in international and comparative advantage. Academy of Management Review, 13: 559-571.

Kleinberg, J.M. (1999). Authoritatives sources in a hyperlinked environment. Journal of the Association for Computing Machinery, 46 (5), 604-632.

Kuhn, T. (1962). The Structure of Scientific Revolutions. Chicago: Chicago University Press.

Larsson, R., Bengtsson, L., Henriksson, K. & Sparks, J. (1998). The interorganizational learning dilemma: collective knowledge development in strategic alliances, Organization Science, 9, 285-305.

Lane, P.J., Koka, B.R. & Pathak, S. (2006). The reification of absorptive capacity: A critical review and rejuvenation of the construct. Academy of Management Review, 31 (4), 833-863.

Mina, A., Ramlogan, R., Tampubolon, G. & Metcalfe, J.S. (2007). Mapping Evolutionary trajectories: Applications to the growth and transformation of medical knowledge. Research Policy, 36, 789-806.

Moed, H. F. (2005). Citation Analysis in Research Evaluation. Dordrecht: Springer.

Noyons, E.C.M. (1999). Bliometric Mapping as a science policy and research management tool. Thesis Leiden University. Leiden: DSWO Press.

Noyons, E.C.M., & Van Raan A.F.J. (1998). Monitoring scientific developments from a dynamic perspective. Self-organized structuring to map neural network research. Journal of the American Society for Information Science (JASIS) 49 (1), 68-81.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web (Tech. Rep.) Standford Digital Library Technologies Project.

Pillai, S.U., Suel, T. & Cha, S.H. (2005). The Perron-Frobenius theorem: some of its applications. IEEE Signal Processing Magazine, 22(2):62 – 75.

Ramlogan, R., Mina, A. Tampubolon, G., & Metcalfe, J.S. (2007). Networks of Knowledge: The Distributed Nature of Medical Innovation. Scientometrics, 70 (2), 459-489.

Small, H.G. (1978). Cited documents as concept symbols. Social Studies of Science, 8, 327-340.

Stiglitz, J.E. (1987). Learning to Learn, Localized Learning and Technological Progress, in P. Dasgupta and P. Stoneman, eds. Economic Policy and Technological Performance. Cambridge: Cambridge University Press.

Zahra, S.A. & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. Academy of Management Review, 27: 185-203.

7 Seed journal citation network maps: A method based on network theory

C.M. Calero Medina , T.N. van Leeuwen (published in the Journal of the American Society for Information Science and Technology, 63(6), 1226-1234, 2012)

7.1 Introduction

Structuring science is about identifying fields, subfields, and research themes and relating them to each other. It is necessary because the traditional science classification system is imperfect, especially for highly multidisciplinary environments, and because it helps to assess performance within its proper context.

In recent years there has been an enormous development in the field of information science in applying different techniques to visualize and analyze the growth of specialties, the structure of scientific communities, and the flow of scientific information (Scharnhorst & Thelwall, 2005).

In fact, as Van Raan (2008) pointed out, science can be considered as an ecosystem comprising species (e.g., fields) whose interdependency can be mapped. The mapping of scientific documents is done in many different ways, depending on the techniques and the purpose on the analysis in which the map is going to be used.

Börner, Chen, and Boyack (2003) reviewed the literature of in bibliometric mapping based on the unit of analysis. The unit of analysis can be documents, relevant terms or words, authors, and journals. Documents are used to visualize and map a knowledge domain with different purposes like analysis of the domain (e.g., Small, 1999) or assessing research performance in a policy context (e.g., Noyons, Moed, & Luwel, 1999). The coword maps are used to unravel the cognitive structure of a field (e.g., Calero, Buter, Cabello, & Noyons, 2006). Authors-based maps are used to infer the intellectual structure of a field (e.g., Chen, 1999). Finally a map of journals can be used to obtain a macro view of science (e.g., Bassecoulard & Zitt, 1999) or to show fine distinctions within a discipline (e.g., Leydesdorff, 1994).

More related to the objective and approach presented in this study is the work done by Leydesdorff and colleagues. In recent years, they have presented a methodology to visualize the citation-impact environment of a given journal (Bornmann, Leydesdorff, & Marx, 2007; Leydesdorff, 2007; Zhou & Leydesdorff, 2007). Their approach makes a distinction between the citing and cited dimensions as two different perspectives on a journal's position (Leydesdorff, 2007). Based on the previous work of He and Pao (1986) and Leydesdorff (1986), the relevant environment for each seed journal (journal under study) is determined by including all journals that cite or are cited by the seed journal to the extent of 1% of its citation rate in the respective dimension. These authors chose the cosine between two vectors (Salton & McGill, 1983) as the similarity measure

between the distributions for the various journals included on the citation environment. Visualization is based on social network analysis techniques.

But in understanding and going a step further in the development of visual maps, we can apply network theory. Scientific documents are interconnected trough citations and coauthorships. The seminal work of Derek de Solla Price (1965) showed the structure of science as a network of interconnected publications. We can explore network structures with the help of complex network theory. In recent years researchers, mainly physicists, have started to use the principles of statistical mechanics to analyze large net-worked structures, including science itself (Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002; Newman, Barabási, & Watts, 2006); thus, network techniques are gradually being applied more intensively in bibliometric analysis. Mapping-interrelated entities enables the study of the topology of complex networks. In science such entities are publications, citations (Menczer, 2004; Van Raan, 2005), journals (Bergstrom, West, & Wiseman, 2008), institutes, and authors (Börner, Maru, & Goldstone, 2004).

7.2 Objectives

Traditional quantitative bibliometric indicators are the standard choice nowadays for assessing the research output of a researcher, research group, and research organization (Moed, de Bruin, & van Leeuwen, 1995). But we have to consider that researchers, historians of science, journal editors, librarians, and science managers are also interested in "larger scale questions" that require assessing hundreds or thousands of research papers by a similar number of authors.

From the perspective of bibliometrics and, particularly, journal performance, our goal with journal-citation network analysis is to be able to provide a quick overview of relevant journals related to a journal under study ("seed journal"), in terms of citations given and received. First, it needs to be established what these journals are, how important they might be, and which position they occupy in the network.

As a starting point, we focus on a specific journal, which is considered the seed journal. This seed journal will have citation links with other journals, both given and received (citing to and cited by). When we have a set of journals, we are able to determine the connections between them based on the citations they give and receive. After that we will extract the most prominent journals using a centrality algorithm developed by

Kleinberg (1999) to separate web pages into authorities and hubs. In our analysis, an important authority journal (with high authority centrality weight) is an important source of scientific knowledge in a given set of journals. An important hub journal (with a high hub centrality weight) is an important source of information to look for the most important authority journals. A journal can have both a high hub centrality weight and a high authority centrality weight at the same time: an important source of scientific knowledge and important source of information to look for the most important authority journals.

Finally, we create a network map that comprises the most important hubs and authorities journals related to the seed journal. In just one network map, we will get the relevant citation environment of a specific seed journal. This approach is new because it considers at the same time the citing and cited dimension of a given journal and uses an algorithm developed in complex network theory to detect the prominent journals. These journal citation network graphs are useful for the various stakeholders in and around the science system, as they provide information on the level of journal connections, unlike the more traditional structures these people are familiar with, such as the Journal Subject Categories, the classification system applied in the products of Thomson Reuters (Journal Citation Reports, Web of Science [WoS], etc.). These network graphs clearly show the closest relations journals can have, based on citation relations, suggesting influence relations between journals in such a way that traditional field boundaries are transcended.

7.3 Methods

In this section we present the data and methods used for this study.

Data

In this study, we start from our CWTS in-house database derived from WoS versions of the Science Citation Index and associated citation indices: the Science Citation Index (SCI), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI). We took all source publications ("articles," "letters," and "reviews") for 2006. The large dataset we created comprises journal-to-journal-citation relations, and it is extracted from the WoS in such a way that all the citing relations of 2006 publications are aggregated to journal level. This means that we grouped the references of (i.e., citations from) the publications in the 2006 source journals to other source publications in the

WoS for the period 1981–2006. So, a citing relation between two journals means a reference (citation given) in 2006 to any other "earlier" publication in the WoS covering the 1981–2006 period. These citations were given by 8,524 journals, citing 8,511 journals covered in the 1981–2006 period. On these data available for the journal citation network analysis, we performed two limiting actions. First we limited the period of citation relations available in the analysis from 1981–2006 to 1997–2006. We monitor then a specific journal (seed journal) and its relations to other journals during a 10-year period (1997–2006). A second limiting action was the creation of symmetry in the dataset, by taking out those journals that are only cited, and not citing journals.

The first step, limiting the period of analysis available in the journal citation network analysis has the following consequences. Initially, this dataset comprised 2,289,383 journal-to-journal relations based on a total of 21,648,745 citations. Limiting the data from the full period to the 10-year period resulted in a dataset of 1,916,714 journal-to-journal relations, and in total 15,528,891 citations. In general, most citations are accounted for within a 10-year window, but it is important to mention that especially for journals in the social sciences and humanities, this limitation is cutting off a larger share of their total number of citations as compared with journals in the natural, life, and technical sciences (Nederhof, 2006). So, the limitation of the period to the 10 most recent years from the year 2006 perspective leads to a loss of 25% of the citations, related to the 1981–1996 period.

Because we want to work with the same journals on the cited range as we work with on the citing range, we limited ourselves to the citing perspective (as this covers all source publications from 2006). In practice, this means that if a journal appeared as a "cited journal" only, we eliminated it from the set. If we then take the next limiting step, the creation of a square matrix of journal-to-journal citation relationships over the citing and cited dimension, for the period 1997–2006, we then started with 8,507 possible journals from which 16 of them appear only as cited by. We removed them from the dataset. This reduction scarcely influences the analysis. This means that at the end what we have is a dataset based on a square matrix (8,491x8,491) of journal-to-journal citation relationships over the citing and cited dimension of 8,491 journals.

Overall, we have asymmetry in the datasets, which comprises two different aspects: first, the asymmetry in the time perspective: the citing year is 2006, cited years are 1997–2006; the second asymmetry comprises the citing-cited relations itself, because a journal can be cited

by another, but this does not have to be the other way around. This creates an asymmetrical matrix, in which the upper part is filled differently than the lower part.

"Seed Journal" citation network

For each seed journal we create a matrix that comprises the journal itself, the journals receiving citations from and giving citations to the seed journal, and all citation connections between these other journals. This is what we called the Seed Journal Citation Network. In terms of network theory, the Seed Journal Citation Network is an "ego network." An ego network comprises a focal node ("ego") and the nodes to which the ego node is directly connected to (these are called "alters") plus the ties, if any, among the alters (Hanneman & Riddle, 2005). In the Seed Journal Citation Network, the nodes are individual journals and the edges are values according to how frequently articles published in one journal (positioned in a row) cite articles published in another journal (positioned in a column). The citations are directional (edges) because a citation from journal B to journal A differs from a citation from A to B (this is the asymmetry of the matrix mentioned above). But there are limitations of using only absolute numbers of citations. In particular, they do not reflect the fact that each number on a cell of the seed journal citation matrix depends on the total number of citations given to and received by the two journals. Thus, we developed an index to measure the relationship between two pairs of journals that controls this bias.

Journal Relationship Measure, L index.

Journal citation rates have been used since the seventies to classify journals and delineate specialty fields (Narin, Carpenter, & Berlt, 1972; Narin & Carpenter, 1973; Leydesdorff, 1994; Narin, Hamilton, & Olivasto, 2000; Pudovkin, 1993; Pudovkin & Fuseler, 1995; Pudovkin & Garfield, 2002). However, none of theses approaches consider at the same time the citing and the cited dimension. The approach we present below takes both into account.

The L index reflects the two dimensions of the matrix "citing" and "cited" and considers the global position of the journals in the Web of Science in terms of total citations given and received. Let C_{BA} be the total number of citations given by journal B to Journal A (or what is the same, the total number of citations received by Journal A from Journal B). Let 'TcitingB' be the total number of citations given by Journal B (in the Web of Science) in 2006 and let be 'TcitedA' the total number of citations received by Journal A between 1997-2006. The L index is

$$L_{BA} = \frac{C_{BA}}{\sqrt{TcitingB * TcitedA}}$$

The L index weights the citations given and received. The citations given from one journal to another are weighted by the total number of citations given by that journal and the total number of citations received by that journal. The L index takes values in the interval [0,1]. It is undefined if the total number of citations given by Journal B or received by Journal A is 0. When the number of citations given by Journal B to Journal A is zero, then the measure is 0. The Index reaches its maximum value of 1, when C_{BA}=TcitingB=TcitedA.

Hubs and Authorities

In network theory, a specific research theme focuses on the identification of important nodes in networks. Garfield's impact factor (Garfield, 1972) is a ranking measure based on counting of in-degrees nodes in a journal citation network. Later, Pinski and Narin (1976) and Geller (1978) developed an algorithm that considered not only the number of citations from one journal to the other but also the prestige of the citing journal. Journals that receive many citations from prestigious journals are considered highly prestigious themselves. By iteratively passing prestige from one journal to the other, a stable solution is reached that reflects the relative prestige of journals (Bollen, Rodriguez, & van de Sompel, 2006). This way of measuring prestige is behind the PageRank algorithms to evaluate the status of web pages, first developed by the founders of the Google Search Engine, Brin and Page (Brin & Page, 1998; Page, Brin, Motwani, & Winograd, 1998). The PageRank is calculated through an iterative algorithm that propagates prestige values from one web page to another and converges to a solution (Pillai, Suel, & Cha, 2005).

At the same time Brin and Page created their Google Search Engine, Kleinberg (1999) constructed an algorithm to increase the effectiveness of Web search engines using the concepts of hubs and authorities. Hubs & Authorities are formal notions of structural prominence of vertices in directed graphs (Brandes & Willhalm, 2002). Following Newman (2010), the centrality algorithm developed by Kleinberg is based on the idea that: "there are really two types of important nodes in a directed network: authorities are nodes that contain useful information on a topic of interest; hubs are nodes that tell us where the best authorities are to be found. (page 179)". An authoritative journal, in our case, is one that is cited by many other journals. This idea can be reinforced by observing that citations from all journals aren't equally valuable – some journals are better hubs (citing journals) for a given journal. The algorithm gives each

node in a network an authority centrality weight and a hub centrality weight. For each journal (j) in a seed citation network we computed two weights: hub centrality weight (hj) and authority centrality weight (aj). The weights show the strength of a given journal as an authority and/or a hub. Weights are computed according to the citation network (M) by solving the eigenvector problem of matrices MM^T (hubs) and M^TM (authorities), where M is the seed citation matrix (Kleinberg 1999). Journal x is considered a more important hub than journal y if $h_x > h_y$. Journal x is considered a more important authority than journal y if $a_x > a_y$. A node with high authority centrality weight is that it is pointed to by many other vertices with high hub centrality weight. And the characteristic of a node with high hub weight is that it points to many nodes with high authority centrality weight (Newman, 2010).

Kleinberg showed examples in which the algorithm could help filter out irrelevant or poor-quality documents (they would have low authority centrality weights) and to identify high-quality documents (they would have high authority centrality weights). Kleinberg (1999) argued that the tradition of the peer review process in scientific journals ensures that the highly authoritative journals with a common purpose reference one another extensively. He considered then that a one-level model (like the one developed by Pinski and Narin, 1976 and Geller, 1978), in which authorities directly endorse other authorities, fits very well. As mentioned above we are analyzing the whole citation environment of a journal, citing and cited dimension together. From our perspective, making a classification in hubs and authorities is a very useful tool to understand the role played by a journal in the citation environment of a seed journal. An important authority journal (with high authority centrality weight) is an important source of scientific knowledge in a given set of journals. An important hub journal (with a high hub centrality weight) is an important source of information to look for the most important authority journals. A journal can have both a high hub centrality weight and a high authority centrality weight at the same time: an important source of scientific knowledge and important source of information to look for the most important authority journals. This is the reason why we decided to use Kleinberg's algorithm for identifying the main journals in the seed citation network. Batagelj adapted for the software Pajek[1] the Kleinberg's hubs/authorities algorithm (Batagelj & Mrvar, 2006). The results from the analysis presented in this article are based on Pajek.

[1] Pajek is a program for Windows, for analysis and visualization of large networks. It was developed by Vladimir Batagelj and Andrej Mrvar. Some procedures were contributed also by Matjaž Zaveršnik.

7.4 Results

To show the results of our method, we have chosen four journals: Scientometrics, Physical Review Letters, Journal of Vascular and Interventional Radiology, and Public Health. The first journal, Scientometrics, is concerned with the quantitative features and characteristics of science. Emphasis is placed on investigations in which the development and mechanism of science are studied by statistical mathematical methods. The second journal selected, Physical Review Letters, is one of the world's foremost physics journals, providing rapid publication of short reports of significant basic research in all fields of physics. International in scope, this journal provides its diverse readership with weekly coverage of major advances in physics and cross-disciplinary developments. The third journal, Journal of Vascular and Interventional Radiology, is the official journal of the Society of Interventional Radiology. Radiologists, cardiologists, vascular surgeons, neurosurgeons, and other clinicians who need current and reliable information on every aspect of vascular and interventional radiology use it. Each issue covers the most critical medical, minimally invasive, radiological, pathological, and socioeconomic issues of importance to vascular and interventional radiologists. The last journal selected is Public Health, a journal aiming at all public health practitioners and researchers and those who manage public health services and systems.

As was described in the previous section, first we selected the journals in the citation environment for each of the four journals analyzed. The selection is based on the journals receiving citations from and giving citations to the seed journal. Table 1 shows the number of journals selected for each of the four journals. The differences among the four journals analyzed already show characteristics of each of these journals. Scientometrics has 271 journals in its citation environment, showing that it is a very specialized journal in certain types of analyses and data. On the other side, Physical Review Letters has 979 journals, showing that it is a general journal in a broad field like physics.

Table 1. Citation environment for each journal

"Seed journal"	*Number journals (citation environment)*
Scientometrics	271
Physical Review Letters	979
Journals of Vascular and Interventional Radiology	443
Public Health	433

The next step was to create, for each of the four journals analyzed, a network that contained the journal itself, the journals receiving citations from and giving citations to the seed journal, and all citation connections between these journals. This is what we called the Seed Journal Citation Network. For instance, the Scientometrics Citation Network is a network of 272 journals (nodes) connected by the absolute number of citations given (or received) from one journal to another. But as we have argued in the previous section, the absolute number of citations is size affected. To avoid it the links between journals are normalized based on the L index described in the previous section. Table 2 shows the minimum and maximum value of the L index in each of the four networks.

Table 2. L index values for each Seed journal citation network

"Seed journal" citation network	**L index (min value)**	**L index (max value)**
Scientometrics	0.0001	0.1410
Physical Review Letters	0.0001	0.2358
Journals of Vascular and Interventional Radiology	0.0001	0.1886
Public Health	0.0001	0.1578

Once the seed journal citation network was normalized based on the L index, we measured the importance of each of the journals in the network using a centrality algorithm developed by Kleinberg (1999) and explained above. The algorithm gave for each journal of the network two weights: authority weight and hub weight. The journals could then be sorted based on these two weights. The journals with the highest weights were selected for being shown in the network map. The decision as to how many journals are selected is arbitrary. We can show in the map as many journals as we want from the seed journal ego network. When we work with the Netdraw program for the visualization of the maps, we can always zoom in or out to get a better view of the journals involved.

Because this is not possible when you make a fixed "image" of the map, we have just selected a "reasonable" amount of journals having the highest weights values based on the hubs and authorities algorithm. In the selection, a journal that has one of the highest hub centrality weights can have also one of the highest centrality weights among the journals in the seed journal citation network. This is a journal considered as an important source of scientific knowledge and an important source of information to look for the most important authority journals among the journals in the citation environment of the given journal.

The maps then show three types of nodes with different shapes. The squares (blue) are the journals with the highest authority weights in the seed journal citation network, the circles (yellow) are the hubs with the highest hubs weights in the seed journal citation network, and the triangles (red) represent the journals that happen to be at the same time in both of the previous selection. The lines (directed edges) show the citation relation between the journals. The direction of the arrow indicates if a journal is cited by (incoming arrow) or if is citing to (outgoing arrow). The thickness of the connecting line reflects the strength of the L index among a pair of journals.

The position of the journals in the map is based in a spring-embedded algorithm included in the software NetDraw. Its effect is to distribute the vertices in a two-dimensional plane with some separation, while attempting to keep connected journals reasonably close together. As de Nooy, Mrvar, and Batagelj (2005) explained, the edges could be imagined as springs "pulling" vertices (journals) together, though never too close. The algorithm pulls vertices to better positions until they reach a state of equilibrium. In the network journal maps, this layout means that journals that are linked or that have links in common will be closer in the map. It is important to consider though that all the journals are appearing on the map because they have been cited by the seed journal. But in the map, we are considering the strongest citation links (based on the L Index) between the journals selected (25% of the links in the map are taken into account). The program used for visualizing the network maps is NetDraw (Borgatti, 2002).

Scientometrics

Figure 1 shows how Scientometrics is between two groups of journals. One is related to information science and technology journals (right-upper part of the network map) and the other with journals related to research, development, and innovation studies, especially from the management perspective (right part of the network map). It is striking to

notice this clear gap between, on the one hand, the scientometrics/library and information science community

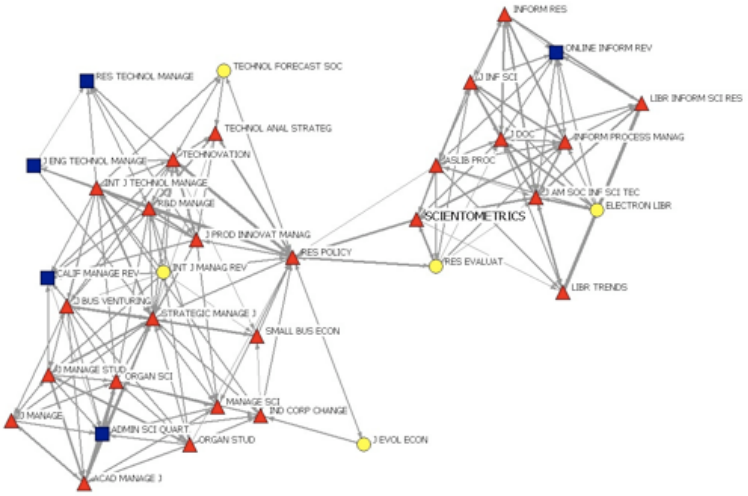

Figure 1. Mapping of the citation environment of Scientometrics (2006) (L index>0.0163)

Squares (blue) Journals with the highest authority centrality weights
Circles(yellow)Journals with the highest hub centrality weights
Triangles(red)-Journals that have at the same time the highest authority and hub centrality weights

Physical Review Letters

Figure 2 shows the central position of Physical Review Letters as well as its status as a hub and authority. Physical Review Letters is first and foremost surrounded by three 'general' or multidisciplinary journals. Phtysical Review B is a general physics journals, while Nature and Science are general science journals. Around this first lay, we notice in the network map also the broad coverage of this journal given its strong connection with journals related with physical sub-disciplines such as astrophysics; elementary particles and fields; nuclear physics; atomic, molecular, and optical physics; nonlinear dynamics, fluid dynamics, classical optics; plasma and beam physics; condensed matter; and soft-matter, biological, and interdisciplinary physics.

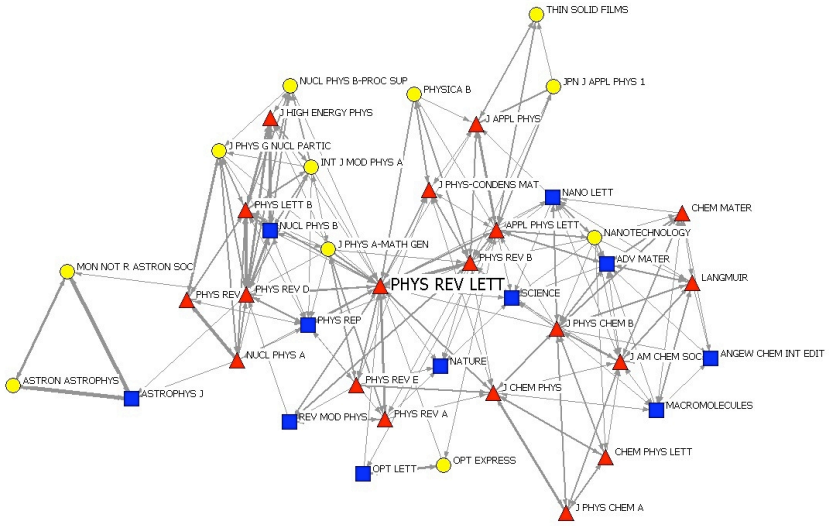

Figure 2. Mapping of the citation environment of Physics Review Letters (2006)
(L index>0.0131)

Squares (blue) Journals with the highest authority centrality weights-
Circles(yellow) Journals with the highest hub centrality weights
Triangles(red)-Journals that have at the same time the highest authority and hub centrality weights

Journal of Vascular and Interventional Radiology (JVIR)

Figure 3 shows the position of this journal between journals related with cardiology and vascular surgery in the upper right side of the graph, and journals related with radiology, neurosurgery, and urology on the other side of the graph. JVIR clearly appears as a hub and authority journal.

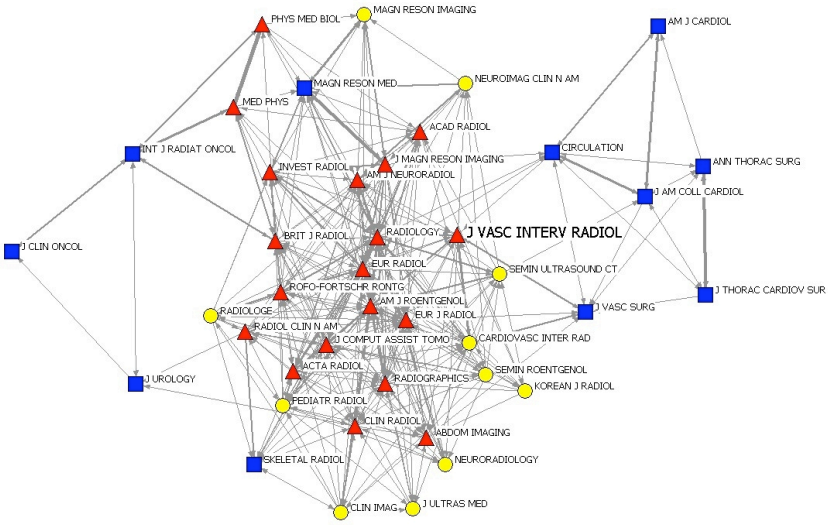

Figure 3. Mapping of the the citation environment of JVIR (2006)
(L index>0.0050)

Squares (blue) Journals with the highest authority centrality weights
Circles(yellow) Journals with the highest hub centrality weights
Triangles(red)-Journals that have at the same time the highest authority and hub centrality weights

Public Health

Public Health is a journal aiming at all public health practitioners and researchers and those who manage public health services and systems. Figure 4 shows its citation network map. Public Health is surrounded by other journals related with public health but none of them have a central position. It is considered a hub spreading the knowledge from journals that are about public sanitary problems as: drugs and addictions, sexual transmittable diseases, obesity, mental health, epidemics, health law and policy.

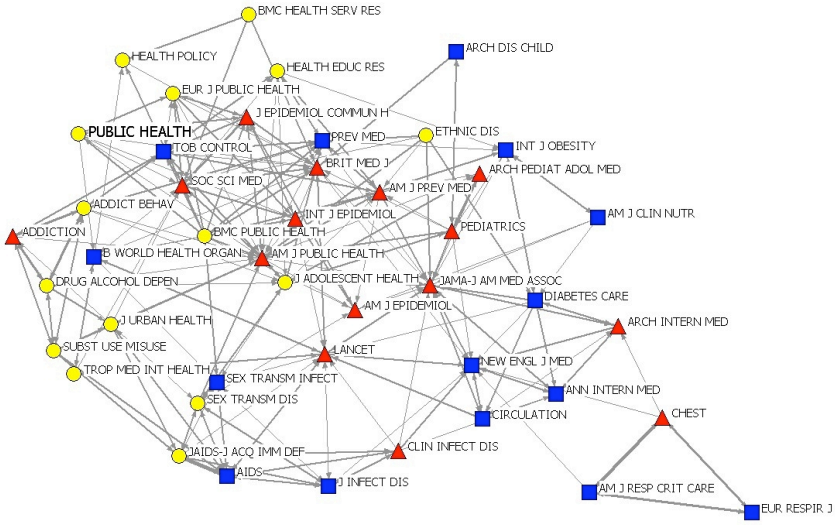

Figure 4. Mapping of the citation environment of Public Health (2006) (L index>0.0089)

Squares (blue) Journals with the highest authority centrality weights
Circles(yellow)Journals with the highest hub centrality weights
Triangles(red)-Journals that have at the same time the highest authority and hub centrality weights

7.5 Conclusions and Follow-Up Research

The method presented here should be considered as the starting point toward a complete methodology for analyzing the citation environment of a journal. From the perspective of bibliometrics and journal performance, our goal was to be able to provide a quick but nevertheless comprehensive overview of the most important journals for a given journal, in terms of citations given and received. The method presented here allows us in a few steps to find the prominent journals, based on a centrality network measure, related to a seed journal. First, we selected the journals in the citation environment of the seed journal. Second, we normalized the citation links inside the seed journal citation network based on an index that considers the importance of the number of citations given from one journal to another in relation to the total number of citations given for the journal and the total number of citations received by the other journal. Then, we apply a centrality algorithm to

determine the importance of the journals in the seed journal citation network. From there on we decide how many of these journals we want to represent in a network map. The decision as to how many journals from the seed journal citation network to include in the map is quite arbitrary though.

We are currently working on a further development of the method presented here through dynamic animation of the network map based on time series of the seed journal network data. The objective is a better understanding of the development through time of the seed journal based on its citation relations. Furthermore, we intend to go a step further in measuring the composition and structure of the seed journal citation network. We are interested in studying how bibliometrically related journals form and evolve embedded in the dynamic system of the seed journal citation network. Measures like homophily (Scott, 2000; Wellman, 1993) can help us to determine if journals that have common bibliometric characteristics (such as journal impact measures, degree of international cooperation, degree of journal-to-journal self citations, etc.) stick together. The study of this phenomenon has also been called "assortative mixing in networks" (Newman, 2002), in which the probability of two nodes being connected by an edge depends on specific similarity properties of the nodes. Another measure called homogeneity can determine whether the seed journal's alters are all alike. We can also analyze the structure of the seed journal (the journals to which the seed journal is connected to) citation network with measures like brokerage and density, which measure whether the seed journal connects otherwise unconnected journals.

In summary, the method and results presented here should be considered a starting point for developing a comprehensive methodology to identify from a dynamic perspective the citation environment of a journal.

References

Albert, R., & Barabási, A.L. (2002). Statistical mechanics of complex networks. Review of Modern Physics, 74, 47–97.

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of journals. Scientometrics, 44, 323-345.

Batagelj, V., & Mrvar, A. (2006). Pajek: Program Package for Large Network Analysis, University of Ljubljana, Slovenia. Retrieved via: http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Bergstrom, C.T., West, J.D., & Wiseman, M.A. (2008). The eigenfactor metrics. Journal of Neurosciences, 28(45),11433-11434.

Bollen, J., Rodriguez, M.A., & Van de Sompel, H. (2006). Journal status. Scientometrics, 69 (3), 669-687.

Borgatti, S. P., (2002). Netdraw: Network Visualization Softwres. Harvard: Analytic Technologies. Retrieved via: http://www.analytictech.com/downloadnd.htm

Börner, K., Chen, C., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science, 37, 179–255.

Börner, K., Maru, J.T., & Goldstone, R.L. (2004). The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences of the USA, 101, 5266–5273.

Bornmann, L., Leydesdorff, L., & Marx, W. (2007). Citation environment of Angewandte Chemie. CHIMIA, 61(3), 104–109.

Brandes, U., & Willhalm, T. (2002). Visualization of bibliographic networks with a reshaped landscape metaphor. Joint Eurographics-IEEE TCVG Symposium on Visualization, D. Ebert, P. Brunet, I. Navazo (Editors). Retrieved via: http://algo.fmi.uni-passau.de/~brandes/publications/bw-vbnr1-02.pdf

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30, 107-117.

Calero, C., Buter, R., Cabello, C., and Noyons E. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. Scientometrics, 66(2), 365–376.

Chen, C.(1999). Visualising semantic spaces and author co-citation networks in digital libraries. Information Processing & Management, 35, 401-420.

De Nooy, W., Mrvar, A. & Batagelj, V. (2005). Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press.

De Solla Price, D.J. (1965). Networks of scientific papers. Science, 149(3683), 510-515.

Dorogovtsev, S.N., & Mendes, J.F.F. (2002). Evolution of networks. Advances in Physics, 51, 1079–1187.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178, 471-479.

Geller, N.L. (1978). On the Citation Influence Methodology of Pinski and Narin. Information Processing & Management, 14, 93-95.

Hanneman, R.A. & Riddle, M. (2005). Introduction to social network methods. Riverside, CA: University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/)

He, C., & Pao, M.L. (1986). A discipline-specific journal selection algorithm. Information Processing & Management, 22(5), 405–416.

Kleinberg, J.M. (1999). Authoritatives sources in a hyperlinked environment. Journal of the Association for Computing Machinery, 46 (5), 604-632.

Leydesdorff, L. (1986). The development of frames of references. Scientometrics,

9(3-4), 103–125.

Leydesdorf, L. (1994). The generation of aggregated journal-journal citation maps on the basis of the CD-ROM version of the Science Citation Index. Scientometrics, 31 (1), 59–84.

Leydesdorff, L. (2007). Visualization of the Citation Impact Environments of cientific Journals: An Online Mapping Exercise. Journal of the American Society for Information Science and Technology, 58 (1), 25-38.

Menczer, F. (2004). Correlated topologies in citation networks and the Web. The European Physical Journal, B38, 211–221.

Moed, H.F., de Bruin, R.E., & van Leeuwen, T.N. (1995). New bibliometric tools for the assessment of National Research Performance—Database description, overview of indicators and first applications, Scientometrics, 33 (3), 381–422.

Narin F. & Carpenter, M.P. (1973). Clustering of scientific journals. Journal of the American Society for Information Science and Technology, 24, 425–435.

Narin, F., Carpenter, M.P., & Berlt, N. (1972). Interrelationships of scientific journals. Journal of the American Society for Information Science and Technology, 23, 323–331.

Narin F., Hamilton, K.S., & Olivasto, D. (2000). The development of science indicators in the United States. In: B. Cronin & H. B. Atkins (Eds). The

Web of Knowledge: A Festschrift in Honor of Eugene Garfield (pp. 337–360). Medford, NJ: Information Today.

Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. Scientometrics, 66, 81-100.

Newman, M.E.J. (2002). Assortative mixing in networks. Physical. Review Letters 89, article no. 208701.

Newman, M.E.J. (2010). Networks: An Introduction. Oxford University Press.

Newman, M.E.J., Barabási, A., and Watts D. (2006). The Structure and Dynamics of Networks. Princeton University Press.

Noyons, E.C.M., Moed, H.F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. Journal of the American Society for Information Science, 50, 115-131.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web (Tech. Rep.) Standford Digital Library Technologies Project.

Pinski, G. & Narin, F. (1976). Citation Influence For Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics. Information Processing & Management, 12, 297-312.

Pillai, S.U., Suel, T., & Cha, S.H. (2005) The Perron-Frobenius theorem: some of its applications. IEEE Signal Processing Magazine, 22(2), 62–75.

Price, D.J. de S. (1965). Network of scientific papers. Science, 149(3683),510-515.

Pudovkin, A.I. (1993). Citation relationships among marine biology journals and those in related fields. Marine Ecology Progress Series, 100, 207-209.

Pudovkin, A.I., & Fuseler, E.A. (1995). Indices of journal citation relatedness and citation relationships among aquatic biology journals. Scientometrics, 32, 227-236.

Pudovkin, A.I., & Garfield, E. (2002). Algorithmic Procedure for Finding Semantically Related Journals. Journal of the American Society for Information Science and Technology, 53(13), 1113-1119.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. Auckland, New Zealand: McGraw-Hill.

Scharnhorst, A., & Thewall, M. (2005). Citation and hyperlink networks. Current Science, 89(9), 1518-1524.

Scott, J. (2000). Social Network Analysis: A Handbook. Newbury Park, CA: Sage Publications.

Small, H. (1999). A passage through science: Crossing disciplinary boundaries. Library Trends, 48, 72-108.

Van Raan, A.F.J. (2005). Reference-based publication networks with episodic memories. Scientometrics, 62(1), 549–566.

Van Raan, A.F.J. (2008). Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of research groups. Journal of the American Society for Information Science and Technology, 59 (4), 565-576.

Wellman, B. (1993). An egocentric network tale. Social Networks 15, 423-436.

Zhou, P. & Leydesdorff, L. (2007). The citation impacts and citation environments of Chinese journals in mathematics. Scientometrics 72(2), 185–200.

8 Conclusions and future prospects

8.1 Key questions

This thesis originated from the need to identify groups of related nodes within collaboration and citation networks. In the study of collaboration networks the main goal is to identify existing research groups, potential research groups, or patterns of collaboration. The analysis of citations networks through specific measures and metrics, on the other hand, makes it possible to identify main lines of research through the years. Thus, such analyses improve our understanding of the growth and decline of fields, including phenomena such as paradigm shifts and emerging research themes. Network measures and metrics also allow for the identification of important nodes (e.g., journals, articles) embedded in the citation network. We addressed three main questions in this thesis:

Can we identify communities, existing research groups, and potential research groups?

Can we identify main lines of research through the years and the articles that linked them into a research tradition that can be considered the backbone of the field?

Can we identify important nodes that play a key role in the citation networks?

In the next sections we will discuss our findings concerning answers to these questions and the necessary future research.

8.2 Results

In **Chapter 2** we presented a method for identifying research groups and potential research partners in scientific fields. We combined a bibliometric science map based on a co-word network with the analysis of a co-publication network. A first and important result of the study is that we have identified *functional* rather than '*physical'* groups. Following Seglen and Aksness's (2000) definition of a research group: "…a research group assignment based on co-authorship defines functional rather than physical groups, and might include, e.g. authors with whom a group member has collaborated in connection with a short-

term scientific visit. Our group concept is thus somewhat wider and loser than the standard conception of a physically localized research team". The groups were defined over a six-year period, which means that the group members had not necessarily worked together. In addition, identification of the members via the combination of author name and affiliation address made it possible for the same person to belong to more than one group. This was the case, for instance, with a researcher who moved from one organization to another and changed his line of research and as a result belonged to two different groups in the period of analysis. A second significant outcome of the study is the possibility to identify potential research partners. Combination of output similarity relations with co-author relations offers a way to detect groups working in the same areas but not co-publishing. A third important result of our approach was that we were able to deal with the problem of homonymous and synonymous author names[1]. The combination of author and address data in a publication allowed us to handle the homonymous names, while the network analysis made it possible to deal with the second category. The combined data enabled us to assign author names to specific researchers more accurately**.**

In **Chapter 3** we presented the case study of how to use publications data to analyze the organizational structure of a large university hospital. Translational research in a university hospital is deeply embedded within daily work activities; it is not limited to a specific hierarchical or technical entity but widely distributed across the entire organization. Thus, proper management is very important in order to facilitate the research activities. In the past years we have observed considerable advances in the development of methods for finding communities within networks, with a large number of different techniques under development. This study shows how bibliometric analyses can benefit from these developments and complement them, since the case studies provided an insight into what the identified groups mean by, validating the results with the opinions of experts involved. The case study presented in Ch. 3 shows how the combination of bibliometric indicators and collaboration analysis can help research managers of large organizations and university hospitals in particular to understand the way the organization behaves, in order to create the strongest possible research clusters.

[1] Homonymous names are two or more persons with the same author name, while synonymous names are two or more different author names referring to the same person.

Chapter 4 describes the results of an empirical study in which we explored the analytical potential of corporate research articles as a source of empirical information for describing structural patterns within multinational enterprises (MNE) in the bio-pharmaceutical industry worldwide, and to produce quantitative data on those research cooperation relationships at the level of countries and major bio-pharmaceutical firms. Given the overwhelming significance of basic research in the bio-pharmaceutical industry and the large quantity of corporate research papers produced each year, we believe that these publications reflect key characteristics of research cooperation patterns within the industry. The outcome revealed interesting empirical information, not only with respect to the organizational features of corporate research partnerships within and between companies, but also on the geographical distribution of these partnerships. The company-level breakdown of these cooperation patterns also reveals a variety of intra- and extra-firm research linkages, from which three main types of corporate research networks can be derived in terms of the intra-firm distribution of research partnerships: (a) centralized networks, (b) decentralized networks, and (c) gateway networks.

Chapter 5 we described a broad study of bibliometric characteristics of largest 386 universities worldwide in terms of number of publications, and of a (partly overlapping) set of 529 European universities. Rather than presenting a ranking, the study presents a statistical analysis of ranking data, focusing on more *general* patterns. Several aspects were compared: US universities with European universities; countries with a strong concentration of academic research activities in a relatively small amount of universities, with nations showing a more even distribution of research over universities; a ranking of universities based on indicators calculated for all research fields combined, with one compiled for a single field (oncology); general with specialised universities; and rankings based on a single indicator with maps combining social network analysis and a series of indicators. The study highlights important factors that should be taken into account in the interpretation of rankings of research universities based on bibliometric indicators. Moreover, it illustrates policy-relevant research questions that may be addressed in secondary analyses of ranking data. In this way, the study was aimed at contributing to a public information system on research universities.

In **Chapter 6** we followed the 'intellectual track' of a specific research concept, *absorptive capacity* (AC), which had a high rate of diffusion through the fifteen years of analysis. With the bibliometric map further concepts (in terms of field-specific keywords) were found which are often related to theories and models associated with the main concept

(AC). Next, we used two other network-based citation-analysis techniques to find the main papers during these years, i.e., the articles that influenced the research for quite a time, and linked them to a research tradition that is the backbone of the 'Absorptive Capacity Field'. Our results show the potential of this methodology as a tool for unraveling the patterns hidden in a set of publications representing a field. The combination of bibliometric mapping with a detailed analysis of the citation network enables us to follow the influence of the introduction of a new concept in a specific research field.

Finally, in **Chapter 7** we presented a method for analyzing the 'citation environment' of a journal. Based on a bibliometric perspective of journal performance, our goal was to provide a fast but nevertheless comprehensive overview of the most important related journals for a given journal in terms of citations given and received. The method introduced in this chapter enabled us to establish the important journals in the citation environment of a given journal, their degree of importance, and the position they occupy in the network.

8.3 Answers to key questions

The answer to the first key question formulated in Section 8.1 is linked with the studies presented in Chapters 2 to 5. We have identified functional research groups embedded in a field (**Chapter 2**) and embedded in an organization (**Chapter 3**), together with potential research groups in a field (**Chapter 2**). We have found broader communities: groups of universities that collaborate based on geographical proximity (**Chapter 5**), and **(Chapter 4)** patterns of intra-firm and extra-firm collaboration.

Chapter 6 is linked to the second question, since in the study described there we identified a main line of research through the years and linked it to a research tradition that can be considered the backbone of the field.

Also in **Chapters 6** together with **Chapter 7,** we identified important nodes that play key roles in two types of citation networks. Thus, these chapters are related with the third question. In **Chapter 6** we identified papers while in **Chapter 7** we identified journals in the relevant citation networks.

8.4 Future Prospects

In general we can say that the future prospects of research as described in this thesis are strongly connected to the reinforcement of the applicability of quantitative studies of science and technology. This is particularly the case for our understanding of knowledge transfer in science and technology, and of directly related themes such as evaluation of research performance, knowledge diffusion, and growth of fields which may be the new sources for innovation. The study of these issues will benefit from the ongoing advances in measures and models of networked systems (citation-based and related networks in our case). They will contribute to a better understanding of the growth and decline of fields, including phenomena such as paradigm shift, emerging research themes, and the establishment of new institutions. Network analysis based on conceptual linkages will substantially improve the mapping of fields in science and technology, and the identification of emerging R&D themes and their actors.

We intend to keep working on the detailed structural properties of citation and collaboration networks. Many networks are characterized by hubs, i.e., nodes of high degree, see for instance Barabasi & Albert (1999); van Raan (2008). Highly cited publications evidently function as hubs, as they are the expression of the phenomenon of preferential attachment in citation networks (Jeong, Neda, & Barabasi (2003)). Mapping of interrelated entities makes it possible to study the topology of complex networks. In science such entities are publications, citations (van Raan, 2000), journals (Bergstrom, West, & Wiseman (2008)), institutes, and authors (Börner, Maru, & Goldstone, 2004). The search for hidden regularities and mathematical expressions to describe them is important because it may reveal the laws underlying the dynamics of complex networked systems (Leicht, Clarkson, Shedden, & Newman, 2007). Most complex networks are the results of a growth process (van Raan (2000), Newman (2001)). Science is an almost perfect example: a dynamical system that evolves through the addition and deletion of nodes and linkages, i.e., by new publications, their references, and newer publications citing older ones. Finding the dynamic rules that govern growth processes will lead to a better understanding of the resulting macroscopic, static properties of networks. To uncover the structure of network growth a rigorous mathematical model is needed. This may shed more light on problems such as the universality of networked systems, classification of networks, hierarchies, and the emergence of clusters, modules, and communities. The ensemble of modules represents highly interlinked communities (Rosvall & Bergstrom, 2007, 2008). How this modularity emerges is one of the basic questions in the study of network

dynamics. Defining the relevant aggregation levels is important in order to understand the relation between citation networks and the impact of authors, and will enable us to find the life lines of science: what was a real breakthrough in the past? Interactions within and between clusters may change, for instance because of the development of a new, interdisciplinary field and its transformation into a mature and stand-alone discipline (Rosvall & Bergstrom, 2010). It is also important to define measurable quantities to describe interactions between time-dependent processes and static topology in the formation of complex networks. Understanding the regulatory and feedback mechanisms connecting various networks is one of the most ambitious goals in network research. Science offers an ideal target to tackle this problem because of the vast amount of data we have available and the presence of clearly observable quantities.

In line with the work described in this thesis we highlight especially the importance of mechanisms connecting citation- and co-authorship networks for the purpose of investigating the role of groups of researchers in the exchange and transfer of knowledge. As mentioned in the introduction of this thesis, the interconnections between scientific publications (e.g. citations given and received from one paper to another) and inside them (e.g. researchers co-authoring papers) allow us to study the way in which scientists create and share new knowledge. Citation networks of scientific publications can be viewed as composed of hierarchically layered networks. The lower network (basic network) consists of citations between scientific publications. A hierarchical step higher than the basic network is the network of citations between researchers. And a further step higher is the network of citations between research groups. Research groups form a crucial aggregation level because they represent the real work floor of science. The exchange of knowledge of research groups measured through the exchange of citations is part of our future interest.

The above issues are linked to another important problem: the identification and definition a research group. Given the large number of empirical studies conducted by CWTS, we have ample information about organizational structures of research institutions, so we can define a research group within the parent organization. In this case the nodes in the higher (third) network are defined by the organizational data. Thus, research groups form an aggregation of organizationally related publications, which is different from bibliometrically related (e.g., co-author based) publications. In other words, basic elements (nodes) of a lower network may also cluster in another network than their own organizational structure. As far as the bibliometrically related

publications concerns, we can identify research groups by looking at authors in co-publication networks. Thus, the co-publication network itself represents a modular structure of co-author groups (Girvan & Newman, 2002; Newman 2004; Newman & Girvan 2004), not necessarily the same as the organization-based research groups to which the authors are affiliated. This bibliometric network and its modules are often more complex than the formal organization, particularly in interdisciplinary research.

Another important question is the translation of the commonly used bibliometric indicators into 'topological' properties of both the lower, basic network as well as of the higher-level networks. For instance, the number of citations of a group is the in-degree of the group in the higher network of groups, and the h-index is a specific variant of the total number of citations at a specific aggregate level (author, group). The impact of a research group in bibliometric terms is the ratio of the number of citations per publication of the group, and the number of citations per publication for the field(s) in which the group publishes. This field-normalized impact represents as it were the fitness of a group as a node in the higher network. The nominator can be seen as a field-specific property of the higher network that encompasses all research groups in science. However, recent studies have shown that normalized indicators are only mathematically consistent if this normalization is not on an aggregate level, but on the lower basic level. Further research is necessary to understand this in the context of network structures.
Furthermore, we intend to go a step further in our attempts to explain how *bibliometrically* related research groups emerge and evolve within the dynamic system of the entire scientific network. This issue is strongly related to cumulative advantage processes which can be analyzed together with other processes that sociologists have studied and found to be important (Powel et all., 2005). For instance, one of these processes is *homophily* (McPherson and Smith-Lovin, 1987), the process by which people who have common characteristics stick together. The study of this phenomenon has also been called 'assortative mixing in networks' (Newman, 2002; Newman, 2003; Newman and Park, 2003) in which the probability of two nodes being connected by an edge depends on specific similarity properties of the nodes. Another interesting process is what Powel et al (2005) have defined as *following the trend*: the network expansion follows a herd-like behaviour, either in response to external pressures, or through what they called 'imitative behaviour'. Finally, the above authors defined the process of *multiconectivity*, that reflects a preference for variety, for moving in different communities and interaction with heterogeneous partners suggesting a search for novelty.

Undoubtedly the research of complex networked systems will benefit from the vast amount of bibliometric data and from the characteristics of bibliometric constructs such as indicators and maps.

References

Barabasi, A.L. and Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.

Bergstrom, C.T., West, J.D.and Wiseman, M.A. (2008). The Eigenfactor Metrics. Journal of Neuroscience, 28(45), 11433–11434.

Borner, K., Maru, J.T. and Goldstone, R.L (2004). The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences, 101 (Suppl 1), 5266.

Girvan, M. and Newman, M.E.J. (2002).Community structure in social and biological networks. Proceedings of the National Academic of Sciences of the USA 99, 7821-7826.

Jeong, H., Neda, Z. and Barabasi, A.L. (2003). Measuring preferential attachment in evolving networks. EPL (Europhysics Letters), 61(4):567–572.

Leicht, E.A, Clarkson, G., Shedden, K. and Newman, M.E.J. (2007). Large-scale structure of time evolving citation networks. The European Physical Journal B, 59(1), 75–83.

McPherson, M. and Smith-Lovin, L. (1987). Homophily in Voluntary Organizations. American Sociological Review 52, 370–79.

Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. Physical Review E, 64(2), 25102.

Newman, M.E.J. (2002). Assortative mixing in networks. Physical Review Letters 89, article no. 208701.

Newman, M.E.J. (2003). Mixing patterns in networks. Physical Review E 67, article no. 026126.

Newman, M.E.J. (2004). Detecting community structure in networks. The European Physical Journal B, 38(2):321–330, 2004.

Newman, M.E.J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* 69, article no. 026113.

Newman, M.E.J. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E* 68, article no. 036122.

Powel, W., White, D., Koput, K. and Owen-Smith J. (2005). Network dynamics and field Evolution: The growth of interorganizational collaboration in the life sciences. American Journal of Sociology 110, 1132-1205.

van Raan, A.F.J. (2000). On growth, ageing and fractal differentiation of science. Scientometrics, 47(2), 347–362.

van Raan, A.F.J. (2005). Reference-based publication networks with episodic memories. Scientometrics, 63(3), 549–566.

van Raan, A.F.J. (2006). Performance-related differences of bibliometric statistical properties of research groups: cumulative advantages and hierarchically layered networks. Journal of the American Society for Information Science and Technology, 57(14), 1919–1935.

van Raan, A.F.J. (2008). Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of research groups. Journal of the American Society for Information Science and Technology, 59(4):565–576, 2008.

Rosvall, M. and Bergstrom, C.T. (2007). An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences, 104(18), 7327–7331.

Rosvall, M. and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(4), 1118–1123.

Rosvall, M. and Bergstrom, C.T. (2010). Quasi-correspondence analysis on scientometric transaction matrices. PLoS ONE, 5(1):e8694.

Seglen, Per O. and Aksnes Dag W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research, Scientometrics 49 (1), 125-143.

Summary

Introduction

The interconnections between scientific publications (e.g., citations given and received from one paper to another) and inside them (e.g., researchers co-authoring papers) allow us to study by means of network analysis the way in which scientists create and share new knowledge. This is a powerful approach to reveal the conditions behind the successful share and transfer of knowledge.

In network theory terminology, the number of citations given to a paper is the *in-degree* of a paper, being a local property of the citation network. This quantity gives information about the characteristics of the network around the nodes, but it does not help to uncover the highly clustered structure of the scientific network. In order to understand the complexity behind knowledge production process, we also need to study the structure of interconnected publications; otherwise we may in fact be missing important and crucial phenomena. Traditionally, the first approach to analyze the structure underlying a network is to make picture of it. During the last years there has been a rapid development in the field of information science applying different techniques to visualize bibliometric networks. Next to visualization techniques ('mapping'), the structural characteristics of scientific networks can be studied using measures and metrics developed in network theory in recent years. These recent developments in general network theory are very useful to incorporate these measures in the studies of scientific networks with the goal of better understanding the process of knowledge creation and sharing.

This thesis originated from the need to identify groups of related nodes within collaboration and citation networks. In the study of collaboration networks the main goal is to identify existing research groups, potential research groups, and patterns of collaboration. The analysis of citations networks through specific measures and metrics, on the other hand, makes it possible to identify main lines of research through the years. Thus, such analyses improve our understanding of the growth and decline of fields, including phenomena such as paradigm shifts and emerging research themes. Network measures and metrics also allow for the identification of important nodes (e.g., journals, articles) embedded in the citation network. We addressed three main questions in this thesis:

* Can we identify communities, existing research groups, and potential research groups?

* Can we identify main lines of research through the years and the articles linking them into a research tradition that can be considered the backbone of the field?

* Can we identify important nodes that play a key role in the citation networks?

The findings concerning answers to these questions are presented below.

Results

The main results of the thesis are presented in Chapter 2 to Chapter 7. In Chapter 2 we present a method for identifying research groups and potential research partners in scientific fields. We combine a bibliometric science map based on a co-word network with the analysis of a co-publication network. A first and important result of the study is that we have identified *functional* rather than '*physical'* groups. Here we follow Seglen and Aksnes' definition of a research group: "…a research group assignment based on co-authorship defines functional rather than physical groups, and might include for instance authors with whom a group member has collaborated in connection with a short-term scientific visit. Our group concept is thus somewhat wider and loser than the standard conception of a physically localized research team". In our analysis the groups were defined over a six-year period, which means that the group members had not necessarily worked together. In addition, identification of the members via the combination of author name and affiliation address made it possible for the same person to belong to more than one group. This was the case, for instance, with a researcher who moved from one organization to another, changed his line of research and as a result belonged to two different groups in the period of analysis. A second significant outcome of the study is the possibility to identify potential research partners. Combination of output similarity relations with co-author relations offers a way to detect groups working in the same areas but not co-publishing. A third important result of our approach was that we were able to deal with the problem of homonymous and synonymous author names. The combination of author and address data in a publication allowed us to handle the homonymous names, while the network analysis made it possible to deal with the second category. The combined data enabled us to assign author names to specific researchers more accurately.

In Chapter 3 we present the study of how to use publication data to analyze the organizational structure of a large university hospital.

Translational research in a university hospital is deeply embedded within daily research work activities; it is however not limited to a specific hierarchical or technical entity but widely distributed across the entire organization. Thus, proper management is very important in order to facilitate such translational research activities. In the past years we have observed considerable advances in the development of methods for finding communities within networks, with a large number of different techniques under development. This study shows how bibliometric analyses can benefit from these developments and complement them, since the case studies provided an insight into what the identified groups mean by validating the results with the opinions of experts involved. The case study presented in Chapter 3 shows how the combination of bibliometric indicators and collaboration analysis can help research managers of large organizations and university hospitals in particular to understand the way the organization behaves, in order to create the strongest possible research clusters.

Chapter 4 describes the results of an empirical study in which we explored the analytical potential of corporate research articles (1) as a source of empirical information for describing structural patterns within multinational enterprises (MNE) in the bio-pharmaceutical industry worldwide; and (2) to produce quantitative data on research cooperation of major bio-pharmaceutical firms at the level of countries. Given the overwhelming significance of basic research in the bio-pharmaceutical industry and the large quantity of corporate research papers produced each year, we believe that these publications reflect key characteristics of research cooperation patterns within this industrial sector. The outcome revealed important empirical information, not only with respect to the organizational features of corporate research partnerships within and between companies, but also on the geographical distribution of these partnerships. The company-level breakdown of these cooperation patterns also reveals a variety of intra- and extra-firm research linkages, from which three main types of corporate research networks can be derived in terms of the intra-firm distribution of research partnerships: (a) centralized networks, (b) decentralized networks, and (c) gateway networks.

In Chapter 5 we describe a broad study of bibliometric characteristics of largest 386 universities worldwide in terms of number of publications, and of a (partly overlapping with the 386) set of 529 European universities. Rather than presenting a ranking, the study presents a statistical analysis of ranking data, focusing on more *general* patterns. Several aspects were compared: US universities with European universities; countries with a strong concentration of academic research

activities in a relatively small amount of universities, with nations showing a more even distribution of research over universities; a ranking of universities based on indicators calculated for all research fields combined, with one compiled for a single field (oncology); general with specialised universities; and rankings based on a single indicator with maps combining social network analysis and a series of indicators. The study highlights important factors that should be taken into account in the interpretation of rankings of research universities based on bibliometric indicators. Moreover, it illustrates policy-relevant research questions that may be addressed in secondary analyses of ranking data.

In Chapter 6 we follow the 'intellectual track' of a specific concept in the field of Organization Research, *absorptive capacity* (AC), which had a high rate of diffusion through the fifteen years of analysis. With the bibliometric mapping method further concepts (in terms of field-specific keywords) were found which are often related to theories and models associated with the main concept (AC). Next, we used two other network-based citation-analysis techniques to find the main papers during these years, i.e., the articles that influenced the research for quite a time, and linked them to a research tradition that is the backbone of the 'Absorptive Capacity Field'. Our results show the potential of this methodology as a tool for unraveling the patterns hidden in a large set of publications representing a field. The combination of bibliometric mapping with detailed analysis of the citation network enables us to follow the influence of the introduction of a new concept in a specific research field.

Finally, in Chapter 7 we present a method for analyzing the 'citation environment' of a journal. Based on a bibliometric perspective of journal performance, our goal was to provide a fast but nevertheless comprehensive overview of the most important related journals for a given journal in terms of citation relations. The method introduced in this chapter enabled us to establish the important journals in the citation environment of a given journal, their degree of importance, and the position they occupy in the network.

Returning to our three key questions formulated in the beginning of this summary this thesis provides the answers to these questions as follows.

The answer to the first key question is linked with the studies presented in Chapters 2 to 5. We have identified functional research groups within a field (Chapter 2) and within an organization (Chapter 3), together with potential research groups in a field (Chapter 2). We also have found broader communities: groups of universities which collaborate on the

basis of geographical proximity (Chapter 5), and (Chapter 4) patterns of intra-firm and extra-firm collaboration.

Chapter 6 is linked to the second question, since in the study described there we identified the main line of research through the years and linked it to a research tradition that can be considered the backbone of the field.

Also in Chapter 6 together with Chapter 7, we identified important nodes that play key roles in two types of citation networks. Thus, these chapters are related with the third question. In Chapter 6 we identified papers while in Chapter 7 we identified journals in the relevant citation networks.

In Chapter 8 the main conclusions and possible future research themes are discussed.

Samenvatting

Inleiding

De relaties die bestaan tussen en binnen wetenschappelijk publicaties (bijvoorbeeld: (citeerrelaties en co-auteurschappen) maken het mogelijk om door middel van netwerkanalyse te bestuderen hoe onderzoekers nieuwe kennis ontwikkelen en met elkaar delen. Deze bibliometrische netwerkanalyse blijkt een krachtig instrument voor het vinden van de voorwaarden waaronder kennis succesvol overgedragen kan worden.

In het jargon van de netwerktheorie noemen we het aantal ontvangen citaties de 'in-degree' van een publicatie, een lokale eigenschap van een citatienetwerk. Het levert informatie over de karakteristieken van het netwerk rond een knoop, maar het zegt nog niets over gehele structuur van het wetenschappelijke netwerk. Om het gecompliceerde proces van kennisproductie te begrijpen, moeten we de gehele structuur van het netwerk te onderzoeken. Op die manier komen we op het spoor van belangrijke zaken. De traditionele benadering is het analyseren van een structuur door deze af te beelden. In het recente verleden zien we binnen de informatiewetenschap een snelle ontwikkeling op het gebied van netwerkvisualisatie. Maar naast analyse op basis van visualisatie (mapping), kunnen we de wetenschappelijke netwerken ook bestuderen op basis van parameters uit de netwerktheorie. De recente ontwikkelingen in dit onderzoeksgebied zijn van groot belang bij het bestuderen van wetenschappelijke netwerken om een beter zicht te krijgen op het proces van kennisproductie en kennisoverdracht.

Het onderzoek in dit proefschrift is voortgekomen uit de wens om clusters van verwante knopen in samenwerkingsnetwerken en citatienetwerken te identificeren en hun betekenis te begrijpen. In samenwerkingsnetwerken kunnen we zowel bestaande onderzoeksgroepen, als mogelijk te creëren groepen, en ook algemene samenwerkingspatronen identificeren. In citatienetwerken vinden we de hoofdlijnen van onderzoek zoals dat zich door de jaren heen ontwikkelt. Op deze manier krijgen we meer zicht op de ontwikkeling van onderzoeksgebieden, in het bijzonder het ontstaan, de groei en het verdwijnen ervan, maar ook het samensmelten van gebieden of het verschuiven van paradigma's. Netwerkanalyse maakt het ook mogelijk om belangrijke knopen in het netwerk te identificeren zoals centrale tijdschriften of invloedrijke artikelen. In dit proefschrift staan drie vragen centraal:

1. Kunnen we met netwerkanalyse ‘research communities’, in termen van bestaande onderzoeksgroepen dan wel mogelijke onderzoeksgroepen identificeren?

2. Kunnen we onderzoekslijnen ontdekken door de jaren heen en de centrale artikelen die de ruggengraat van deze ontwikkelingen vormen?
3. Kunnen we de knopen met een centrale rol in een citatienetwerk identificeren?

De bevindingen in mijn onderzoek met betrekking tot deze vragen worden in de hoofdstukken 2-7 gepresenteerd.

Resultaten

In hoofdstuk 2 bespreken we een methode om in een specifiek onderzoeksgebied onderzoeksgroepen te identificeren en mogelijke partners voor samenwerking. Onze benadering betreft een combinatie van een 'co-woord' netwerk (gebaseerd op het samen voorkomen van bepaalde trefwoorden/concepten in publicaties) en een met 'co-publicatie' gegevens. In deze studie worden niet zozeer formele *organisatorische* groepen gevonden maar meer de *functionele* groepen. In dit opzicht volgen we de definitie van onderzoeksgroepen van Seglen en Aksnes: "…a research group assignment based on co-authorship defines functional rather than physical groups, and might include for instance authors with whom a group member has collaborated in connection with a short-term scientific visit. Our group concept is thus somewhat wider and loser than the standard conception of a physically localized research team". In onze analyse worden functionele groepen geïdentificeerd binnen een periode van zes jaar wat betekent dat de betrokken onderzoekers niet noodzakelijkerwijs met elkaar hoeven te hebben samengewerkt. De betrokken onderzoekers (auteurs) worden gevonden door combinatie van naam en werkadres wat betekent dat een persoon gedurende de bestudeerde periode tot meer dan één groep kan behoren. Dit is bijvoorbeeld het geval als een onderzoeker binnen de periode van zes jaar 'verhuist' van de ene organisatie naar de andere en daarbij van onderzoeksthema verandert.

Een tweede resultaat van deze studie is de mogelijkheid om potentiële onderzoekspartners te identificeren. Een combinatie van relaties op basis van verwante onderzoeksthema's en samenwerking (coauteurs) stelt ons in staat onderzoekers te vinden die niet samenwerken maar wel in het zelfde vakgebied actief zijn. Een derde belangrijk resultaat van deze benadering is de aanpak van het probleem rond homonieme (één naam die naar meerdere personen verwijst) en synonieme (meerdere namen die naar één persoon verwijzen) auteursnamen. De combinatie van naam met adres lost het probleem van de homonieme namen grotendeels op, terwijl netwerkanalyse de synonieme namen op kan sporen. Deze benadering

stelt ons in staat om nauwkeurig ‘reële’ personen te koppelen aan namen zoals die voorkomen in gegevensbestanden.

In hoofdstuk 3 gebruiken we publicatiegegevens om de organisatorische structuur van een groot academisch ziekenhuis te analyseren. In medische centra is translationeel onderzoek diep geworteld in de dagelijkse praktijk maar het is niet georganiseerd in specifieke eenheden of op een bepaald niveau binnen de hiërarchie. Veeleer is het verdeeld over de gehele organisatie. Om die reden is goed management van translationeel onderzoek van groot belang. Recentelijk is er belangrijke vooruitgang geboekt in de ontwikkeling van methoden en technieken om specifieke ‘communities’, vooral met betrekking tot translationeel onderzoek, te identificeren binnen bestaande netwerken. In deze studie laten we zien hoe bibliometrische analyses hier een belangrijke bijdrage kunnen leveren en tevens methodologisch verbeterd kunnen worden door validatie van de resultaten van de case studies zijn gevalideerd door experts. De conclusie is dat combinatie van netwerkanalyse en bibliometrische indicatoren gebruikt kan worden bij het monitoren van het onderzoek in een grote organisatie zoals een universitair medisch centrum, en bij het nemen van strategische beslissingen.

In hoofdstuk 4 presenteren we een bibliometrische studie van publicaties van bedrijven. We willen de toepasbaarheid van deze methode aantonen met betrekking tot (1) het beschrijven van structurele patronen binnen het onderzoek bij multinationals in de biofarmaceutische industrie; en (2) het beschikbaar maken van kwantitatieve informatie over onderzoeksamenwerking tussen biofarmaceutische bedrijven op het niveau van landen. Gegeven het sterk toenemende belang van onderzoek binnen de biofarmaceutische industrie en (daarmee samenhangend) de grote hoeveelheid wetenschappelijke publicaties van deze bedrijven, levert bibliometrische analyse van deze publicaties belangrijke informatie over van samenwerkingspatronen binnen het industriële onderzoek in deze sector. Deze empirisch vastgestelde informatie betreft niet alleen de organisatie van samenwerking tussen en binnen bedrijven, maar ook de geografische spreiding van deze samenwerking. Deze samenwerkingsanalyse van onderdelen binnen de grote bedrijven (intern en extern) levert drie type netwerken op: (a) gecentraliseerde netwerken; (b) gedecentraliseerde netwerken; en (c) 'gateway' netwerken.

Hoofdstuk 5 presenteert een brede studie van de bibliometrische karakteristieken van de 386 grootste universiteiten in de wereld en van 529 (deels overlappend met de set van 386)) Europese universiteiten. Deze studie levert een ranking van universiteiten op, maar het belang van de studie is vooral gelegen in de statistische analyse van de ranking,

gericht op het vinden van algemene patronen. Een aantal aspecten wordt onderzocht en vergeleken: het verschil tussen Europese en Amerikaanse universiteiten; verschillen tussen landen met relatief veel en relatief weinig concentratie van onderzoek in een beperkt aantal universiteiten; verschillen tussen universiteiten op basis van alle wetenschapsgebieden samen, en op basis van een specifiek gebied (in dit geval: oncologie); verschillen tussen algemene (brede) en gespecialiseerde universiteiten; en rankings gebaseerd op één indicator versus een netwerkanalyse van de betreffende universiteiten gecombineerd met een reeks performance indicatoren. De studie levert een aantal belangrijke factoren bij de interpretatie van rankings van universiteiten gebaseerd op bibliometrische indicatoren. Bovendien laat deze studie zien hoe secundaire analyse van ranking data relevant is bij de beantwoording van vragen op het gebied van onderzoeksbeleid.

In hoofdstuk 6 volgen we het 'intellectuele pad' van een specifiek onderwerp binnen het vakgebied Organisation Research: *absorptive capacity* (AC). Dit onderwerp kenmerkt zich door een hoge mate van diffusie gedurende de periode van vijftien jaar waarover wij dit proces bestuderen. Met co-woord analyse identificeren wij binnen deze onderzoekslijn belangrijke concepten die gerelateerd kunnen worden aan bestaande theorieën en modellen. Daarnaast passen we twee andere netwerktechnieken toe om binnen dit AC onderzoek de centrale publicaties te vinden. Dit zijn publicaties die de ontwikkeling van het vakgebied gedurende een bepaalde tijd sterk hebben beïnvloed. Deze publicaties hebben we verbonden aan de concepten die kenmerkend zijn voor de betrokken publicaties, en deze combinatie vormt de ‘ruggengraat’ van het betrokken vakgebied. Onze benadering toont de kracht van deze specifieke netwerkanalyse om verborgen structuren binnen een onderzoeksgebied bloot te leggen. De tijdsdimensie voegt een belangrijk element toe: combinatie van bibliometrische co-woord analyse en gedetailleerde analyse van een citatienetwerk stelt ons in staat om de diffusie in de loop der tijd van een nieuw concept binnen een onderzoeksgebied te volgen.

Ten slotte presenteren we in hoofdstuk 7 een methode om de ‘citatieomgeving’ van een tijdschrift te analyseren. De analyse beoogt een snel maar toch een zo volledig mogelijk overzicht van andere tijdschriften die belangrijk zijn (d.w.z. wat betreft gebied aan elkaar verwant) voor een bepaald tijdschrift, op basis van citeerrelaties van en naar elkaar. Met de ontwikkelde methode zijn we in staat om de belangrijkste tijdschriften te identificeren voor een bepaald tijdschrift en hun positie in het netwerk waarvan ze deel uitmaken.

Terugkerend naar onze drie centrale onderzoeksvragen genoemd in het begin van deze samenvatting levert dit proefschrift de beantwoording van deze vragen als volgt.

Hoofdstuk 2 tot en met 5 behandelen vraag 1. In deze hoofdstukken identificeren we functionele onderzoeksgroepen binnen een vakgebied (hoofdstuk 2) en binnen een organisatie (hoofdstuk 3). Verder gaan we in op het identificeren van mogelijke nieuwe onderzoeksgroepen in hoofdstuk 2 In hoofdstuk 5 hebben we bredere 'onderzoekscommunities' geïdentificeerd: groepen van samenwerkende universiteiten gebaseerd op geografische nabijheid. Hoofdstuk 4 beschrijft het vinden van patronen van samenwerkingsrelaties binnen bedrijven en tussen bedrijven en andere onderzoeksinstellingen.

Hoofdstuk 6 gaat in op de tweede vraag waarbij we de ontwikkeling van een vakgebied volgen met behulp van de voorgestelde methode en daarmee de centrale artikelen als 'ruggengraat' van het vakgebied identificeren. Hoofdstuk 6 behandelt tevens, evenals hoofdstuk 7, het vinden van belangrijke 'knopen' in twee typen citatienetwerken (de derde vraag). In hoofdstuk 6 betreffen deze netwerkknopen publicaties, en in hoofdstuk 7 zijn het tijdschriften.

In hoofdstuk 8 worden de belangrijkste bevindingen en voorstellen voor verder onderzoek besproken.

Curriculum Vitae

Clara Maria Calero Medina was born the 25th of April 1973 in Madrid, Spain. After a few weeks she moved with her parents to Andalucía where she lived until the age of 17. Her last year of high school she spent it in the USA where she got her high school diploma in 1991.

Clara studied Economics at Carlos III University in Madrid. Right after she finished her degree in 1996, she got a scholarship to keep her studies in the same University in the specialization of analysis and management of science and technology. During the next year and a half her studies were focused on: strategic analysis of economic sectors, management of innovation, public support and corporate technology strategies, technological change and economic growth, management of international co-operation, comparative analysis of national systems of Innovation, and R&D Evaluation.

Her specialization studies were supported by her work as research assistant for the Laboratory for Analysis and Assessment of Technical Change in Carlos III University between 1996 and 1999. Her work at the Laboratory was mainly related with two big projects. The first one was the evaluation of the PETRI programme (Programme for the Stimulation of the Transfer of Research Results), a national initiative provided by the Spanish National R&D Plan. The criteria for the evaluation was primarily to determine the social, technical, structural and strategic achievements of the programme in order to inform policy-making process on the nature and quality of co-operative projects. The second one was a project financed by the European Union (EU). The objective of the project was to describe the evolution of policies towards Research Joint Ventures in a specified representative sample of seven EU member states and compare them with the respective policies practiced in the United States and Japan.

In 1999 she got a scholarship and went to the École Normale Supérieure (ENS) in Paris. There she assisted to some of the statistics courses from the ENS and other institutions in Paris. Back in Madrid in September 2000 she worked during 6 months as project manager for the General Research Head Office of the Regional Government of Madrid.

After that she went back to her old research group in Carlos III University. She worked in a bibliometric analysis of the publications produced by the Central Bank of Spain since 1984. During this time she was also working as tutor in Economics for some of the degrees in Carlos III University.

In November 2002 she participated with the Centre for Science & Technology Studies (CWTS) at Leiden University in a project financed by the Spanish government where a bibliometric evaluation of Spanish research groups in 7 fields of the 6th EU framework was carried out. Since 2003 she works at CWTS. During these years she has been involved in numerous bibliometric research performance studies mainly under the group of Dr. van Leeuwen. In parallel she carried out her Ph.D. research in the topic of citations and co-authorship networks as a bibliometric tool for science evaluation purposes under the supervision of Professor A.F.J. van Raan.